MOLECULAR EVOLUTIONARY AND POPULATION GENETICS ANALYSES OF HUMAN H1N1 HA AND NA GENES IN PANDEMIC AND NON-PANDEMIC YEARS

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfilment of the Requirements for the Degree of

MASTER OF SCIENCE

in Biotechnology

by
Kıvanç NAYCI

December 2023
İZMİR

We approve the thesis of **Kıvanç NAYCI**

Examining Committee Members:

Assoc. Prof. Dr. Efe SEZGİN
Department of Food Engineering, Izmir Institute of Technology

Prof. Dr. Çağlar Karakaya
Department of Molecular Biology and Genetics, Izmir Institute of Technology

Assoc. Prof. Dr. Zeynep A. KOÇER
Department of Molecular Biology and Genetics, Izmir Biomedicine and Genome Center

14 December 2023

Assoc. Prof. Dr. Efe SEZGİN
Supervisor, Department of Food Engineering
Izmir Institute of Technology

Assoc. Prof. Dr. Ali Oğuz BÜYÜKKİLECİ
Head of the Department of
of Biotechnology

Prof. Dr. Mehtap EANES
Dean of the Graduate School
Engineering and Science

ACKNOWLEDGEMENTS

Firstly, I would like to thank my advisor Assoc. Prof. Efe Sezgin for his guidance, his knowledge, and his teachings. Without him, this thesis would not be possible.

Secondly, I would like to thank Elif Kaplan, my dear friend and labmate who always helped me with my questions and helped me to keep going. My awesome friends, Alex and Marion were always here for me when I needed support the most, and their support was invaluable to me.

Lastly, I would like to thank my family and especially my sister. Their constant motivation was helpful in writing this thesis.

ABSTRACT

MOLECULAR EVOLUTIONARY AND POPULATION GENETICS ANALYSES OF HUMAN H1N1 HA AND NA GENES IN PANDEMIC AND NON-PANDEMIC YEARS

The 1918 H1N1 pandemic, known as Spanish Flu, is one of the deadliest pandemics on recorded history. It is estimated that the Spanish Flu pandemic affected over 500 million people across the globe, and the death toll is estimated to be between 20 to 50 million. Ever since this, scientists worked hard to find an effective vaccine for influenza, but its very rapidly evolving nature made this task quite the challenge. In this thesis we performed molecular evolution and population genetics analyses on 35714 hemagglutinin and 36302 neuraminidase nucleotide sequences to better understand the evolution of these proteins. The Tajima's D values showed strong positive selection on the pandemic year of 2009 and the BEAST analysis results also suggested there was a greater exponential growth compared to other years. The relaxation of selection and lack of exponential population growth was inferred from the calculations for 2021 sequences, whereas the positive selection on the hemagglutinin and neuraminidase proteins was evident for the 2022 sequences. Outgroup tests also confirmed the positive selection was acting on the pandemic and non-pandemic years, the tests also confirmed the divergence of human influenza neuraminidase from the swine influenza neuraminidase. HA2 part of hemagglutinin and 475-500 nt part of neuraminidase proteins were found to be the most conserved parts of these proteins. In conclusion, the positive selection on these two proteins returned after the year 2021, which was a very unusual year for influenza that caused the positive selection on the virus and the exponential growth rates of the virus to decline. The most conserved regions can be a good candidate for small molecule/drug and vaccine studies.

ÖZET

İNSAN H1N1 VİRÜSÜNÜN HA VE NA GENLERİNİN PANDEMİ VE PANDEMİ OLMAYAN YILLARDA MOLEKÜLER EVRİM VE POPÜLASYON GENETİK ANALİZLERİ

İspanyol Gribi olarak bilinen 1918 H1N1 salgını, kayıtlı tarihin en ölümcül salgınlarından biridir. İspanyol Gribi salgınının dünya genelinde 500 milyondan fazla insanı etkilediği, ölü sayısının ise 20 ila 50 milyon arasında olduğu tahmin ediliyor. O tarihten bu yana bilim insanları gribe karşı etkili bir aşı bulmak için çok çalışıyor fakat grip virüsünün çok hızlı değişen doğası, bu görevi oldukça zorlu hale getirdi. Bu tezde, grip virüsünün önemli olan hemagglutinin ve nöraminidaz proteinlerinin evrimini daha iyi anlamak için 35714 hemaglutinin ve 36302 nöraminidaz nükleotid dizileri üzerinde moleküler evrim ve popülasyon genetiği analizleri yaptık. Tajima'nın D değerleri, 2009 pandemi yılında olan güçlü bir pozitif seçilim gösterdi ve bulgular aynı zamanda diğer yıllara nazaran daha fazla bir üstel nüfus artışına işaret etti. 2021 dizileri için seçilimin gevşetildiği ve üstel popülasyon büyümesinin olmadığı sonucu çıkarılırken, hemaglutinin ve nöraminidaz proteinleri üzerindeki pozitif seçim 2022 dizileri için açıktı. Dış grup testleri ayrıca pandemik ve pandemik olmayan yıllardaki pozitif seçimi ve ayrıca insan influenza nöraminidazının domuz influenza nöraminidazından farklılığını doğruladı. Hemaglutinin'in HA2 kısmı ve nöraminidaz proteinlerinin 475-500 nt'lik kısmının bu proteinlerin en çok korunan kısımları olduğu belirlendi. Sonuç olarak, influenza için çok olağandışı durumların yaşandığı, virus üzerindeki pozitif seçilimin ve popülasyon büyümesinin azalmasına neden olan 2021 yılından sonra bu iki protein üzerindeki pozitif seçilim geri döndüğünü gördük. En çok korunan bölgeler küçük molekül/ilaç ve aşı çalışmaları için iyi bir aday olabilir.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

Eta	Total Number of Mutations
HA	Hemagglutinin
HA1	Hemagglutinin subunit 1
HA2	Hemagglutinin subunit 2
IAV	Influenza A
IAB	Influenza B
IAC	Influenza C
IAD	Influenza D
JC	Jukes-Cantor Corrected
M1	Influenza Matrix Protein 1
M2	Influenza Matrix Protein 2
mRNA	Messenger Ribonucleic Acid
NA	Neuraminidase
NP	Influenza Nucleoprotein
NS1	Influenza Nonstructural Protein 1
NS2	Influenza Nonstructural Protein 2
PA	Viral Polymerase Subunit
PB1	Viral Polymerase Subunit 1
PB2	Viral Polymerase Subunit 2
pdm09-like	2009 pandemic strain H1N1 influenza virus
QIV	Quadrivalent Influenza Vaccine
RNA	Ribonucleic Acid

S	Segregating Sites
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
TD	Tajima’s D test
TIV	Trivalent Influenza Vaccine
2009pdm-like	2009 pandemic strain H1N1 influenza virus
π	Nucleotide Diversity
*	$P < 0.05$
**	$P < 0.01$
***	$P < 0.0001$
#	$0.10 < P < 0.05$

CHAPTER 1

INTRODUCTION

1.1 Influenza Classification

Influenza viruses belong to the group of negative-sense, single-stranded RNA viruses. They have the ability to infect a wide range of hosts, including avians, swine, cattle, and humans, leading to the well-known respiratory illness called "flu." Their capacity to cross species barriers makes them a constant concern for public health, requiring ongoing monitoring and research [1].

Viral particles of influenza can reach between 80-120nm in diameter. Influenza viruses belong to the *orthomyxoviridae* family of viruses [2].

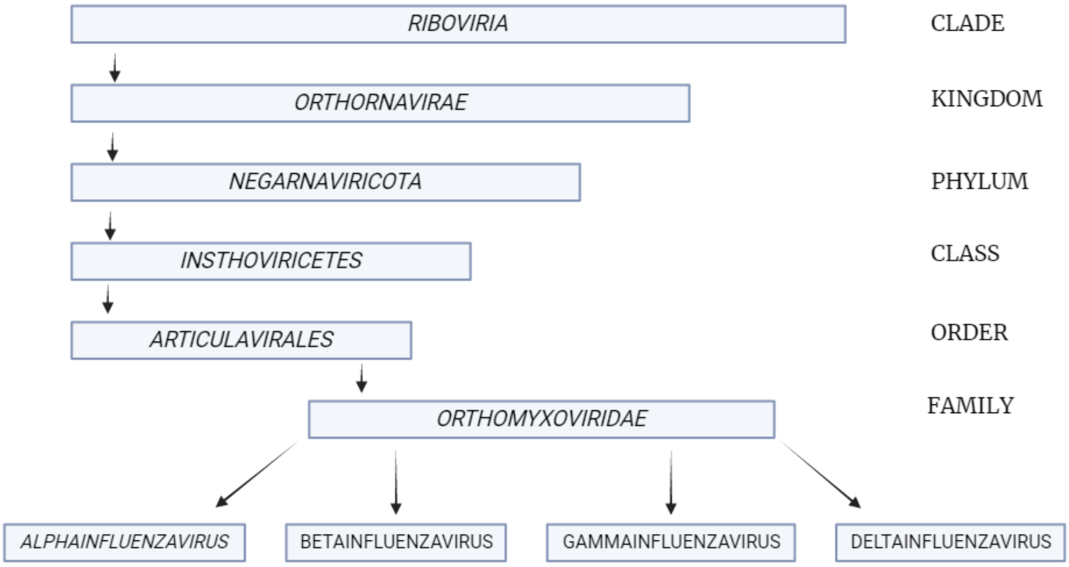


Figure 1.1 Classification of Influenza viruses. (Made with Biorender)

There are four recognized types of influenza: influenza A (IAV), influenza B (IBV), influenza C (ICV), and influenza D (IDV). Scientists closely study these types to understand their unique characteristics, virulence, and patterns of transmission, in order to develop effective strategies for prevention and control.

Meanwhile four different subtypes of influenza exist, currently only influenza A and influenza B are known to infect humans. The disease, flu, is an upper respiratory infection caused by the influenza virus. Common symptoms of this disease are fever, coughs, a sore throat, lack of energy (lethargy), stuffed nose. The main way of transmitting this disease is airborne particles from an infected individual [3].

1.2 Structure of Influenza A Virus

Influenza A is a single stranded, negative-sense RNA virus with eight RNA segments. These RNA segments are individually contained in viral ribonucleoproteins, and these eight segments are packaged within the viral envelope to form a virus particle.

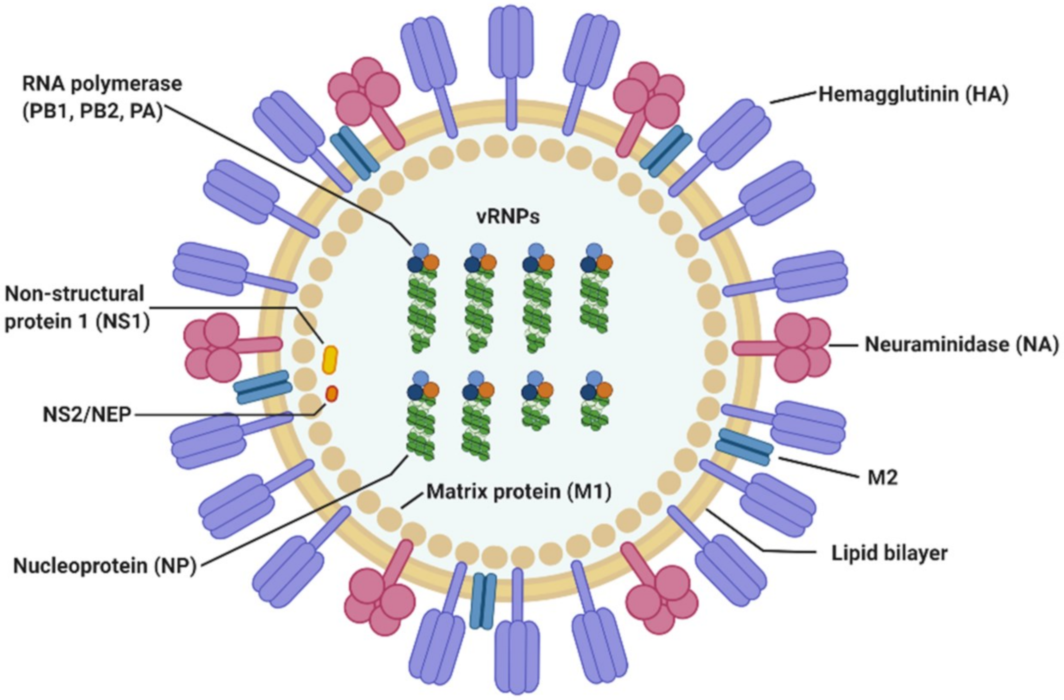


Figure 1.2 The structure of influenza A virus[4].

The eight RNA segments are responsible for producing a total of 11 essential proteins, each serving distinct functions. Scientists have extensively studied the roles of these proteins [4 5 6].

Hemagglutinin (HA) protein plays a crucial part in enabling the virus to enter host cells by facilitating its attachment to the cell surface. This specific role makes it an attractive target for investigations aimed at developing antiviral treatments and vaccines, as blocking HA's action could hinder the infection process [7 8].

Similarly, the neuraminidase (NA) protein is vital for the release of newly formed viral particles from infected cells, aiding the virus in spreading throughout the respiratory tract. There are studies that are trying to develop antiviral drugs that are targeted at inhibiting NA’s functions to limit viral replication [9 10].

Other significant proteins include matrix protein 1 (M1) and matrix protein 2 (M2), which are involved in virus assembly and release, safeguarding the virus's structural integrity and promoting its genetic material's release into host cells. The influenza nucleoprotein (NP) plays a pivotal role in packaging the viral RNA segments and protecting them during replication, contributing to viral RNA synthesis and assembly [4 11].

The non-structural proteins, NS1 and NS2, are multifunctional, influencing various aspects of the virus's life cycle. NS1 counters the host's immune response, while NS2 participates in viral assembly and regulation [11].

The viral polymerase subunits, PA, PB1, and PB2, work in tandem to support viral replication and transcription. They are crucial for creating new viral RNA segments during replication [12].

Lastly, the PB1-F2 protein has been implicated in enhancing the virus's pathogenicity and influencing infection severity. Researchers are conducting numerous studies to comprehend its role and potential as a therapeutic target [12].

1.2.1 Hemagglutinin (HA)

Hemagglutinin is a surface protein on influenza virion that facilitates the entrance to the host cell. HA is a homo-trimetric protein that is glycosylated. HA will recognize and bind to the sialic acid that is on the host cell membrane. This recognition causes a conformation change in HA's structure, in which HA starts the membrane fusion of virus and the host cell; therefore, facilitating the entrance to the host cell [13].

HA is made up of two different subunits called HA_1 and HA_2 respectively. Native HA that has not gone under any conformational change is called HA_0. HA_1 subunit contains the receptor-binding site for sialic acid and is located at the top of the HA protein. HA_2 subunit is bound to the HA_1 subunit with disulfide links, and they are responsible for binding the HA protein to the viral membrane and contains the fusion peptide that is used for fusing viral and host membranes [7 14].

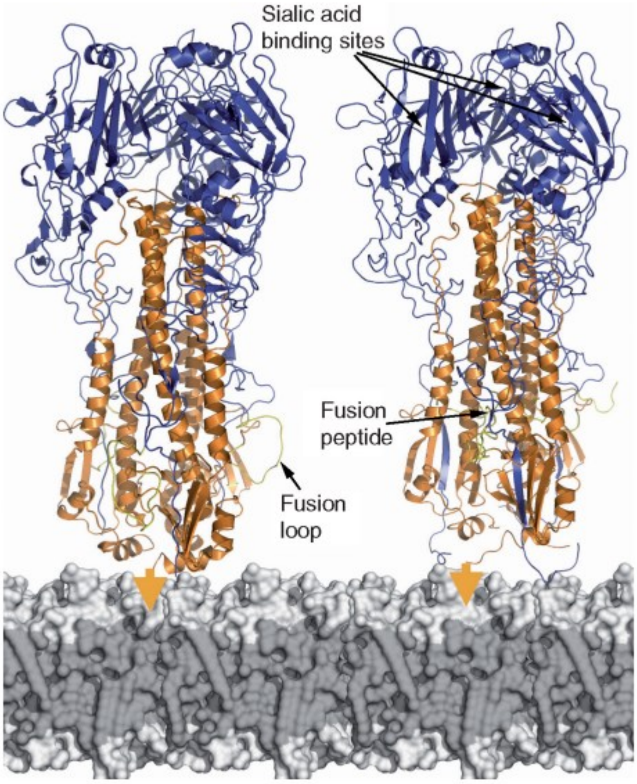


Figure 1.3 Structure of HA protein before the proteolytic processing (left) and after (right). Blue part is the HA1, orange part is the HA2[15].

When HA is in low pH environment, HA_1 will be destabilized, which is required for the activation of fusion peptide that is in HA_2. This conformational change happens when HA_1 receptor-binding site binds to the sialic acid on the host membrane, which allows fusion protein to undergo conformational changes and expose the fusion peptide. This allows the fusion of virus and host cell membranes [15] [16].

1.2.2 Neuraminidase (NA)

Neuraminidase is another surface protein, and it is responsible for facilitating viral release from host cell by cleaving the α-ketosidic link between sialic acid and a nearby sugar molecule, which facilitates the release of newly formed viral particles from the infected host cell. This process is necessary because the newly formed virions use the host cell's cell membrane for their own membrane, meaning the newly synthesized HA and NA molecules will be inserted into the membrane. This causes newly synthesized HA proteins to bind with the host's sialic acid receptors. If these bonds are not cleaved, newly formed viral particles cannot exit the host cell, therefore they cannot infect more cells [17] [18] [19].

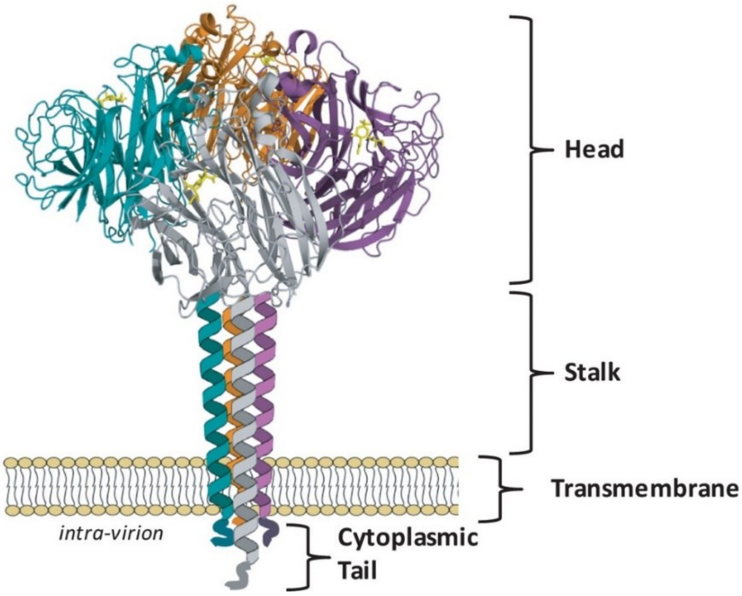
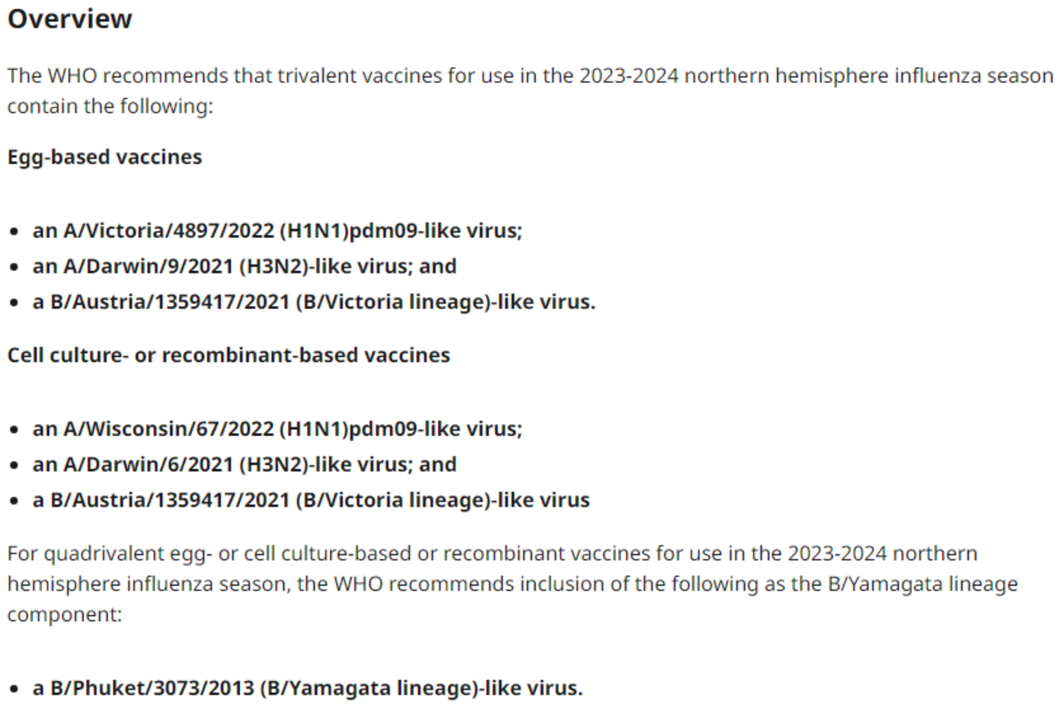


Figure 1.4 Structure of neuraminidase (NA) on influenza membrane. The four parts of the tetramer are all shown in different colors (orange, grey, purple, green)[10].

Studies also showed that other than allowing newly formed viral particles to leave, NA also might have functions that help with infecting the host animal during first contact. NA helps the viral particles to navigate the mucus filled airway of the animal, which is rich with sialic acid receptors. These mucus-rich airways can help the body fight the infection by acting as “decoys” and holding the infecting particles near entrance of the opening where it is easier for body to fight it [17 18 19].

1.3 Current Influenza Vaccine Efforts

CDC (Centers for Disease Control and Prevention) states that the currently used influenza vaccines are usually 40-60% effective when the infecting influenza type and influenza type that used to make the vaccine matches [20]. These vaccines are considered as “old-generation” and every year, WHO (World Health Organisation) selects four candidates of influenza strains to be put in that year’s vaccination composition [21].



Overview

The WHO recommends that trivalent vaccines for use in the 2023-2024 northern hemisphere influenza season contain the following:

Egg-based vaccines

- **an A/Victoria/4897/2022 (H1N1)pdm09-like virus;**
- **an A/Darwin/9/2021 (H3N2)-like virus; and**
- **a B/Austria/1359417/2021 (B/Victoria lineage)-like virus.**

Cell culture- or recombinant-based vaccines

- **an A/Wisconsin/67/2022 (H1N1)pdm09-like virus;**
- **an A/Darwin/6/2021 (H3N2)-like virus; and**
- **a B/Austria/1359417/2021 (B/Victoria lineage)-like virus**

For quadrivalent egg- or cell culture-based or recombinant vaccines for use in the 2023-2024 northern hemisphere influenza season, the WHO recommends inclusion of the following as the B/Yamagata lineage component:

- **a B/Phuket/3073/2013 (B/Yamagata lineage)-like virus.**

Figure 1.5 WHO vaccine composition recommendation for northern hemisphere 2023-2024 season[22].

1.3.1 Classical Influenza Vaccines

1.3.1.1 Egg Based Influenza Vaccines

The classical influenza vaccines are made in two main ways: using fertilized chicken eggs and more recently, using cell cultures. For the egg-made vaccine process, the candidate viruses are picked, and passaged into the fertilized eggs where they will multiply and be ready for harvest. This method does not always work on viruses, and in that case the virus is passaged into the egg with a "donor" virus that is viable in egg. This way, there is a chance that the donor and the unviable virus swap gene segments, resulting in a virus with wanted specific antigenic properties (candidate virus antigenic parts) with egg-viable segments. This way the unviable candidate viruses could be produced for vaccines. After the incubation period, the allantoic fluid is harvested and influenza virus is killed or inactivated with chemicals to be used in vaccines [21 23 24].

The traditional egg-based production method is a well-established approach but comes with certain challenges. In the case of inactivated vaccines, each vaccine dose necessitates three to four chicken eggs, depending on whether it's a trivalent inactivated vaccine (TIV) or a quadrivalent influenza vaccine (QIV). This requires the coordination of the production of over 100 million embryonated chicken eggs from pathogen-free flocks. Maintaining the cleanliness of these flocks and ensuring sterility during the manufacturing process can be difficult. Hygiene lapses can result in the rejection of large quantities of vaccine. Additionally, not all virus strains thrive in embryonated hens' eggs, particularly H3N2 strains [25].

Crucially, the manufacturing process for inactivated influenza vaccines require the influenza virus to infect the cells used (such as avian cells in eggs). Seasonal viruses naturally grow in humans and, as a result, can grow in certain mammalian cells. To infect a cell, the influenza virus must bind to a cellular receptor. Avian cells have different receptors compared to mammalian cells. Therefore, for a human influenza virus to grow effectively in avian cells, it must adapt to bind to the avian receptor, a process known as egg adaptation. Unfortunately, the region on the influenza virus where this adaptation occurs is the same region that is antigenically dominant. This means that as the virus adapts to grow in eggs, it has the potential to differ antigenically from circulating viruses. This difference may lead egg-based vaccines to be potentially less effective in preventing influenza infection compared to their non-egg adapted counterparts grown in mammalian cells [25 26].

1.3.1.2 Cell Culture Based Influenza Vaccines

The cell culture technology in influenza vaccine making eliminates the need to have millions of healthy flock eggs to be used in the vaccine making. It also eliminates the need to use avian-binding influenza virus to be incubated in egg, which resulted in lower antigenicity for human vaccine uses. In this method, the candidate viruses are grown in Madin-Darby Canine Kidney (MDCK) cells, which are mammalian kidney cells. Furthermore, cell-based production is not reliant on a steady supply of eggs. Instead, MDCK cells can be frozen in large quantities if there is a need, especially during a pandemic. Both cell-based and recombinant influenza vaccines are accessible for use in the United States, and recently, a cell-based quadrivalent influenza vaccine has also received approval in the EU. Initiatives are in progress to make these vaccines available in more regions [25 26].

Studies have demonstrated that viruses derived from cells are more closely aligned with circulating viruses compared to those derived from eggs. Analysing data from the WHO Collaborating Centre for influenza, researchers have assessed the similarity between circulating viruses from the 2003 season and those from the 2017–2018 season with selected MDCK-derived and egg-derived viruses, some of which were chosen for the respective season's vaccine [25 26].

1.3.2 mRNA Vaccines for Influenza

Following the SARS-CoV2 pandemic and the efforts for an effective mRNA vaccine for COVID-19, the work for an mRNA based vaccine for influenza picked up the pace. Top pharmaceutical companies are developing mRNA vaccines for influenza which, if successful, could be much more efficient in influenza prevention compared to the classical vaccines. While there are very few vaccines in clinical trials, there are more being developed and will be entering clinical trials[27].

Table 1.1. mRNA vaccines that are being developed[27].

Developer	Product Name	Type	Antigen
In Clinical Trials			
Pfizer	PF-07252220	Monovalent (H1N1) and (B/Yamagata), to be combined into bivalent and quadrivalent	Hemagglutinin
Moderna	mRNA-1010	Quadrivalent	Hemagglutinin
Sanofi/Translate Bio	MRT-5400, MRT-5401	Monovalent (H3N2)	Hemagglutinin
Preclinical State			
Moderna	mRNA-1020 mRNA-1030	Multivalent	Hemagglutinin+ Neuraminidase
Moderna	mRNA-1073	Quadrivalent + COVID-19	Hemagglutinin
Sanofi		Quadrivalent	Hemagglutinin
NIAID		Monovalent (H1N1)	Hemagglutinin
NIAID		Universal	Hemagglutinin stem
Innorna		Quadrivalent and Pentavalent (two H3N2 Strains)	Hemagglutinin

1.4 2009 H1N1 Influenza Pandemic (Swine Flu Pandemic)

The 1918 Spanish flu pandemic witnessed a significant event when the influenza virus jumped from avians to humans, an unprecedented cross-species transmission at that time. This occurrence paved the way for future zoonotic transmissions, impacting global health for years to come. Almost a century later, in 2009, the swine flu pandemic emerged, with the virus jumping from swine to humans, causing widespread concern worldwide [28 29 30 31].

The 2009 swine flu was yet another deadly pandemic caused by an influenza virus, with first being the 1918-1920 Spanish flu and the second one the 1977 Russian flu pandemics for H1N1 [32 33]. The 2009 pandemic influenza was first discovered in two distinct cases in United States during April of 2009 [34]. The 2009 pandemic H1N1 was a new strain formed from a triple assortment of avian, swine, and human influenzas which then further combined with Eurasian swine flu, earning its name the “swine flu” [35 36].

There were more than 50 thousand confirmed cases worldwide, however this number is believed to be severely underestimated. In a study it is estimated that the real number of cases were ranging between 700 million to 1.4 billion people, including mild symptomatic cases and asymptomatic cases. This equates to 11 to 21 percent of world population at the pandemic years[37].

The 2009 swine flu pandemic served as a wake-up call, prompting worldwide efforts to improve preparedness and collaboration, ensuring a safer and healthier world for future generations[38].

1.5 Hypothesis and Aims of the Study

The hypothesis of this thesis is studying the molecular evolution of 2009 pandemic H1N1 influenza viral sequences can give valuable information about the types of selection acting on the hemagglutinin and neuraminidase proteins of the virus. These findings can be used to identify the most conserved and fastest evolving parts of these proteins that can be utilized for better vaccine and drug (small molecule) development and monitor global viral dynamics.

To address this hypothesis firstly, the H1N1 HA and NA nucleotide sequences from the 2009 pandemic, 2015-2021 non-pandemic, and 2022 SARS-CoV2 post-pandemic years from six continents are gathered and their phylogenetic topologies are compared in a Bayesian time-tree framework. Secondly, molecular population genetics and selection analyses compared and contrasted the dynamics of molecular evolution in different years and in different continents. The nature of selection (positive, negative or balancing) acting on HA and NA sequences are inferred.

CHAPTER 2

MATERIALS AND METHODS

2.1 Data Collection and Preparation

The HA and NA sequences for the data analyses were all collected from GISAID (https://gisaid.org)[39]. Only fully sequenced data was collected, in order to be able to analyse the full nucleotide sequences of HA and NA proteins. However, in order for Dnasp software to use the data, there could be no unambiguous characters in the files [40]. To remove the sequences that contained unambiguous characters, firstly all unambiguous characters were turned into "N" using USA government's official HIV sequence database's format conversion tool (https://www.hiv.lanl.gov/content/sequence/FORMAT_CONVERSION/form.html). After this conversion, all the sequences that contained the letter "N" were removed by using a Biopython script. After removing those sequences, all the sequences were aligned using MAFFT 7.4872.1 alignment tool[41].

Table 2.1. The number of HA nucleotide sequences for each year and continent.

	Africa	Asia	Australia	Europe	N. America	S. America
2009	124	941	140	727	1937	66
2015	60	781	42	598	427	51
2016	271	770	163	1315	1714	439
2017	315	922	137	394	895	43
2018	350	1369	399	1489	2209	364
2019	129	1641	501	2124	3555	65
2020	179	453	129	1382	2117	84
2021	162	18	-	9	17	-
2022	258	355	238	1936	765	118

Table 2.2. The number of NA nucleotide sequences for each year and continent.

	Africa	Asia	Australia	Europe	N. America	S. America
2009	131	980	140	771	2090	64
2015	73	893	40	641	427	96
2016	279	891	154	1401	1714	75
2017	346	1015	140	389	895	46
2018	378	1597	80	1604	2355	75
2019	137	1964	514	2281	3555	80
2020	198	565	140	1455	2117	94
2021	170	10	-	16	17	-
2022	215	336	243	1657	670	88

The BioPython script for clearing out the sequences that contain the letter “N” is as follows[42]:

```
“import sys
from Bio import SeqIO
for record in SeqIO.parse(sys.argv[1], "fasta"):
  if record.seq.count('N') == 0:
    print(record.format("fasta"))”
```

The outgroups for HA sequences were picked as human H5 sequences, with the following GISAID isolate IDs: EPI_ISL_379574, EPI_ISL_195732, EPI_ISL_195659, EPI_ISL_8799552, EPI_ISL_305763. For NA, swine NA1 sequences were used as outgroup. The GISAID isolate IDs for the swine NA outgroups are: EPI_ISL_238917, EPI_ISL_281883, EPI_ISL_304066, EPI_ISL_304068.

2.2 Population Genetics and Molecular Evolution Analyses

All the population genetics and molecular evolution analyses were done by using DnaSP6 software[40]. Analyses were done using the aligned nucleotide sequences of HA and NA. The tests for population genetics analyses are: calculation of nucleotide diversity, Theta-W (Watterson theta estimation) calculation, and neutrality tests: Tajima’s D, Fu-Li’s D, Fu-Li’s F, Fu-Li’s D*, Fu-Li’s F*. In addition to these tests that tests for neutrality, McDonald-Kreitman (MK) test and Fay and Wu's H tests were used with outgroups to test the direction of selection in the sequences, as well as direction of selection test[43]. The Greek letter π, stands for nucleotide diversity in the analyses and it shows the average number of nucleotide differences between two sequences in the alignment [44]. Watterson estimator, θ, is used to show the nucleotide proportion distribution of polymorphic sites [45].

In addition to these parameters, neutrality tests give very valuable information about the selection on the sequences by calculating neutrality. Tajima's D shows the selection that is on the population and how is the population size changing. If the Tajima's D value is zero, it signifies that there is no indication of selection upon the population. A positive Tajima's D value signifies that there is a negative selection on the population, and the size of population is decreasing. A negative Tajima's D value shows that there is a positive selection for a specific allele, and the population is expanding in a quick fashion [46]. Another measurement of selection on the population is the Fu-Li's tests. For Fu-Li's tests, a negative number means that there is a great number of singletons in the sample while a positive number means the samples lack a good number of singletons. The tests, Fu-Li's D* and Fu-Li's F* are used to measure the singletons interspecies method (only the same species of data) meanwhile Fu-Li's D and Fu-Li's F measurements require an outgroup to perform singleton calculations[47].

Other calculations for selection on the populations are Fay and Wu's Hn test, and McDonald-Kreitman tests. Fay and Wu's Hn test measures the selection on the population based on a change in population size or a selective sweep in the population [48]. The McDonald–Kreitman test serves as a statistical tool frequently employed by evolutionary and population biologists. Its primary purpose is to identify and quantify adaptive evolution occurring within a species. This test achieves this by assessing whether adaptive evolution has taken place and by determining the proportion of genetic substitutions resulting from positive selection, also referred to as directional selection. In the context of the McDonald-Kreitman (MK) Test, negative selection is indicated when the neutrality index is less than one. This happens when the rate of genetic changes that lead to amino acid differences (nonsynonymous) between species is lower than the rate of such changes within a single species. On the other hand, positive selection is observed when the neutrality index exceeds one. This occurs when there is a higher proportion of explained genetic variations in the opposite direction. Additionally, the test calculates a P-value based on the Chi-square value to assess the statistical significance of these findings [49] [50] [51].

2.3 BEAST Analyses for Evolution Simulations

BEAST (Bayesian Evolutionary Analysis Sampling Trees) is a program that uses MCMC (Markov chain Monte Carlo) for Bayesian analysis [52]. To use BEAST, firstly a BEAST XML file was made by using its companion software, BEAUti. The aligned nucleotide sequences were loaded into BEAUti 1.10.4, where it is possible to select origins year of sequences, the molecular clock model [53], nucleotide substitution model, population growth mode. For this thesis, all the sequences were analyzed with GTR substitution model, with 4 gamma categories for site heterogeneity, and nucleotide positions unlinked into three partitions. The clock model is random local clock [53], and the tree prior is exponential growth, which is selected by looking at Tajima's D scores. All files were run for 1 million Markov chains, logging parameters every 1000 Markov chain. The outputs of BEAST 1.10.4 analyses were read using Tracer 1.7.2 [54].

2.4 Tree Construction and Visualization

The consensus tree for each of the analyses was made by using the BEAST companion software TreeAnnotator 1.10.4. Maximum clade credibility trees were constructed using the tree files that was created by BEAST software after the Bayesian MCMC analyses alongside with log files. The constructed tree files were visualized with FigTree 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) software [55] and TreeViewer software [56](https://treeviewer.org). The colorings for the trees were made with a BioPython script.

CHAPTER 3

RESULTS AND DISCUSSION

In total 35714 HA and 36302 NA sequences were used for the analyses in this thesis. North America, Europe, and Asia continents contributed to most of the number of sequences meanwhile South America continent contributed the least. For 2021, at the time of gathering sequences and conducting analyses, there were no sequences on GISAID flu database for South America and Australia continent. The specific numbers of sequences per year for each continent can be found in Tables 2.1 for HA, and Table 2.2 for NA.

3.1 Timetree Phylogenetic Analyses of 2009pdm H1N1 HA and NA Sequences

The timetrees of HA and NA were constructed using sequences from all years across all the continents. The selection was based on the percentage of each sample contributes to the total number of samples. If the number corresponded to the percentage was lower than 5, it was counted as 5.

The HA tree topology shows a distinct separation during 2017, with both clades seemingly cascade into many smaller clades that go on throughout the years. There are no data for years 2010-2014, therefore it is not possible to comment on the divergence patterns in those years. Comparing 2009 and rest of the tree, the sequences are much closer to each other; a drastic separation cannot be observed like there is in 2017, or in 2018.

The continents which are the closest relative to the clade that leads to the next year is not constant. The closest relative of 2015 clade is North America, meanwhile the closest relative of 2016 clades are Africa and Europe. The sequences are also not clustered together based on their continent of origins. These findings are in line with the GISAID flu genomic epidemiology HA timetree (appendix).

While some years have very distinct separations such as 2015, 2017, 2018 other years do not have a distinct separation, like 2009, 2016, 2020. This is further supported by the posterior probabilities of the branches of the tree (Supplementary Figure 55).

The posterior probabilities for the separation of 2017 branches are low, however it is the best tree topology out of ten thousand (Supplementary Figure 55).

Table 3.1. Number of sequences used to construct the 2009 pandemic HA timetree with all years and continent included (2994 sequences used in total).

	Africa	Asia	Australia	Europe	N. America	S. America
2009	10	79	11	61	162	5
2015	5	65	5	50	35	5
2016	33	64	13	110	143	36
2017	26	77	11	33	75	5
2018	29	114	33	125	185	30
2019	10	137	42	178	298	5
2020	15	38	10	116	177	7
2021	13	5	5	0	5	0
2022	21	29	19	164	64	9

The NA tree showed clearer distinction for each years, with each year showing up as at least two distinct clades, with separate branches leading to the clades. Similar to HA tree, the 2009 data is closer to each other compared to the other years. Because there are no 2010-2014 sequence data it is not possible to claim which continent is closer to the 2015 clade. Similar to HA tree, the continents do not group within each other and the sequences belonging to different continents can be seen as closer to each other rather than being closer to the other sequences from the same continents.

The posterior probabilities for separation of year’s branches are not high, however the tree topology is the best out of ten thousand tree tries (Supplementary Figure 56).

Similarly, the NA tree does not have a specific continent in which the next year’s clades are formed from. This is further supported by the GISAID flu genomic epidemiology NA timetree (Supplementary Figure 54).

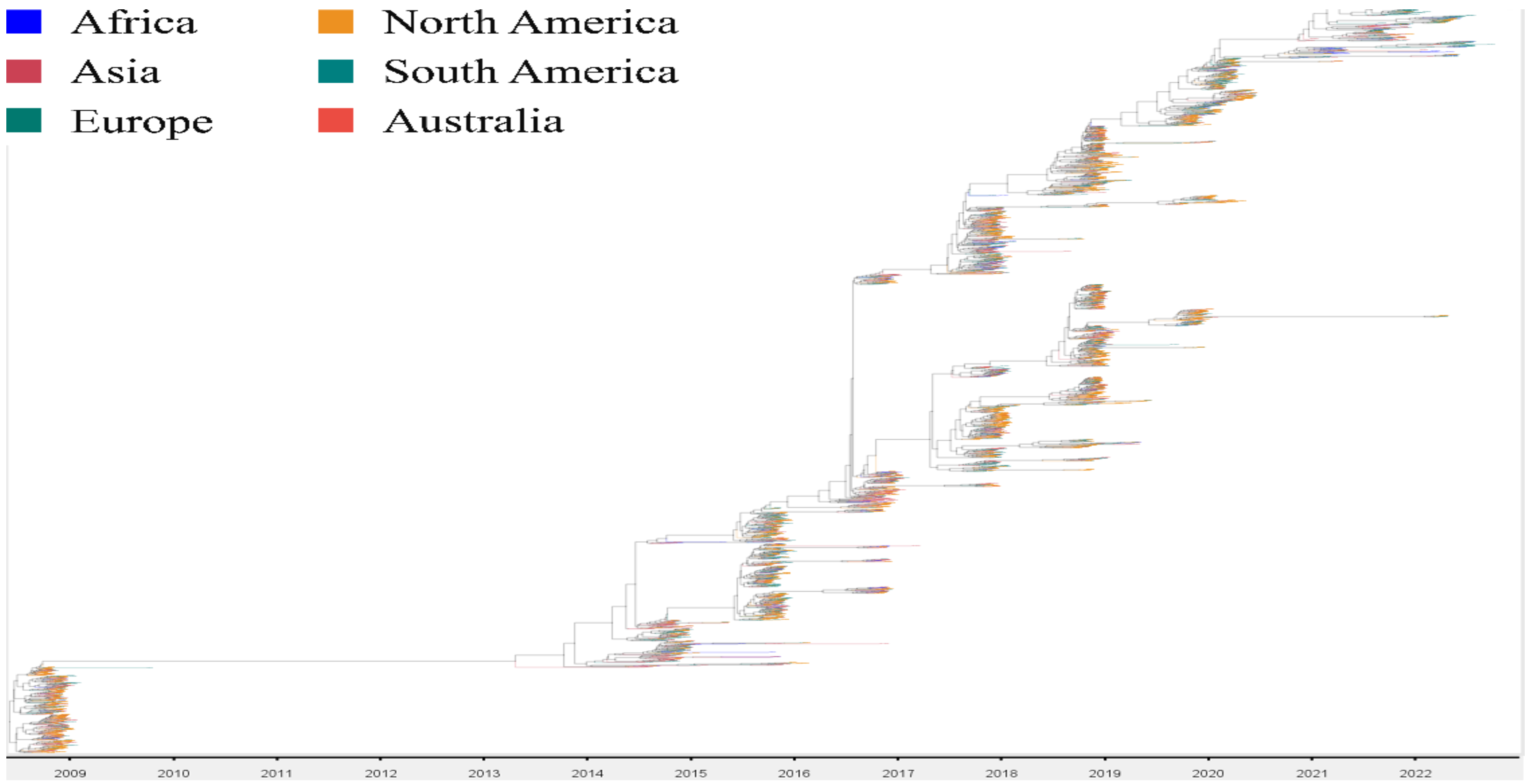


Figure 3.1 The 2009 pandemic H1N1 time tree showing the phylogenetic relationship between HA nucleotide sequences from six continents and nine years (2009 – 2022).

Table 3.2 Number of sequences used to construct the 2009 pandemic NA timetree with all years and continent included (2989 sequences used in total).

	Africa	Asia	Australia	Europe	N. America	S. America
2009	10	80	10	63	172	5
2015	6	73	5	52	35	7
2016	23	73	12	115	141	6
2017	28	83	11	32	73	5
2018	31	131	6	132	194	5
2019	11	162	42	188	293	6
2020	16	46	11	120	174	7
2021	14	5	0	5	5	0
2022	17	27	20	136	55	7

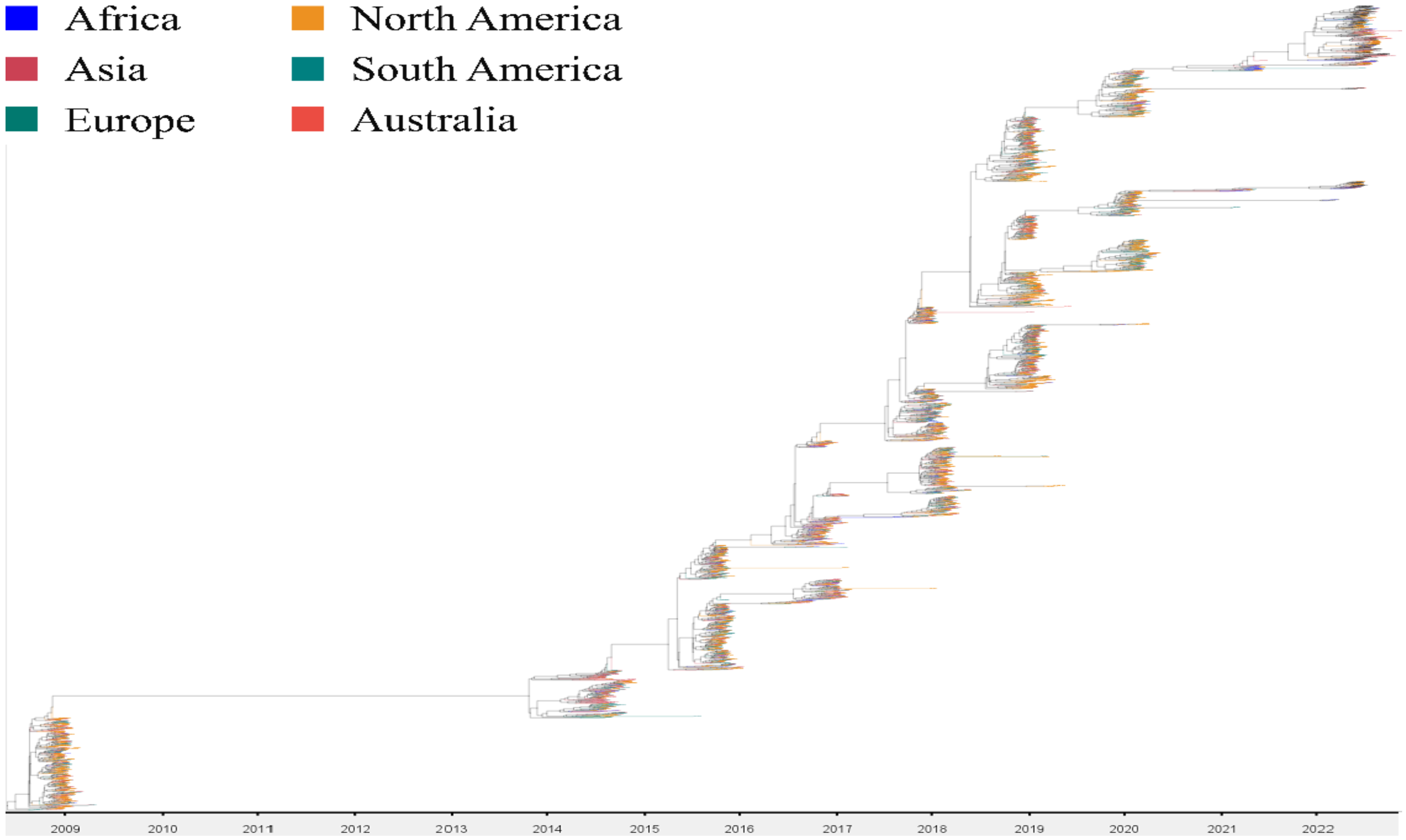


Figure 3.2 The 2009 pandemic H1N1 time tree showing the phylogenetic relationship between NA nucleotide sequences from six continents and nine years (2009 – 2022).

The tree topology is similar to the HA tree (Figure 3.1), however there are no distinct clades for 2017 and 2018. It can also be seen that the 2022 sequences are closer to each other compared to other years on the tree. It is seen that some years have ancestors back more than two years, indicating that these viruses still might be circulating in the population. For some of the sequences, branch lengths of 2022 is longer than other years. The low amount of sequences for year 2021 could be a reason for this.

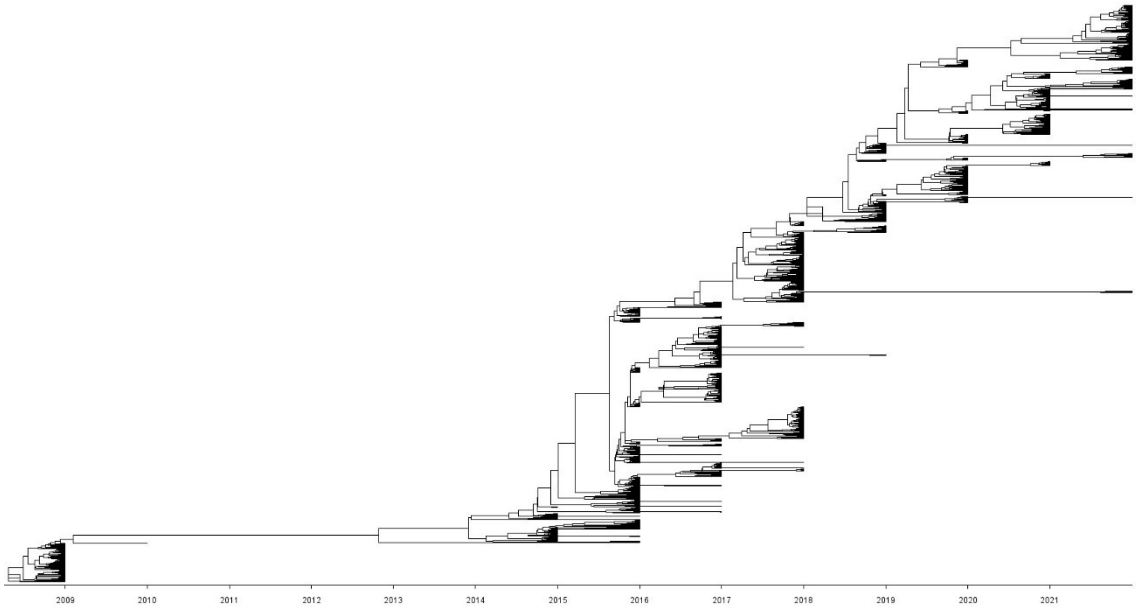


Figure 3.3 The 2009 pandemic H1N1 time tree showing the phylogenetic relationship between HA nucleotide sequences from Africa continent across nine years (2009 – 2022).

The topology of the tree is similar to the NA tree (Figure 3.2). Compared to the HA tree of Africa continent, the 2022 sequences are closer to each other, similar to those of 2009. The branch lengths of 2022 is similar to other years, except one sample. This shows that for these samples, the lineage of NA is not interrupted by a lack of infections in 2021.

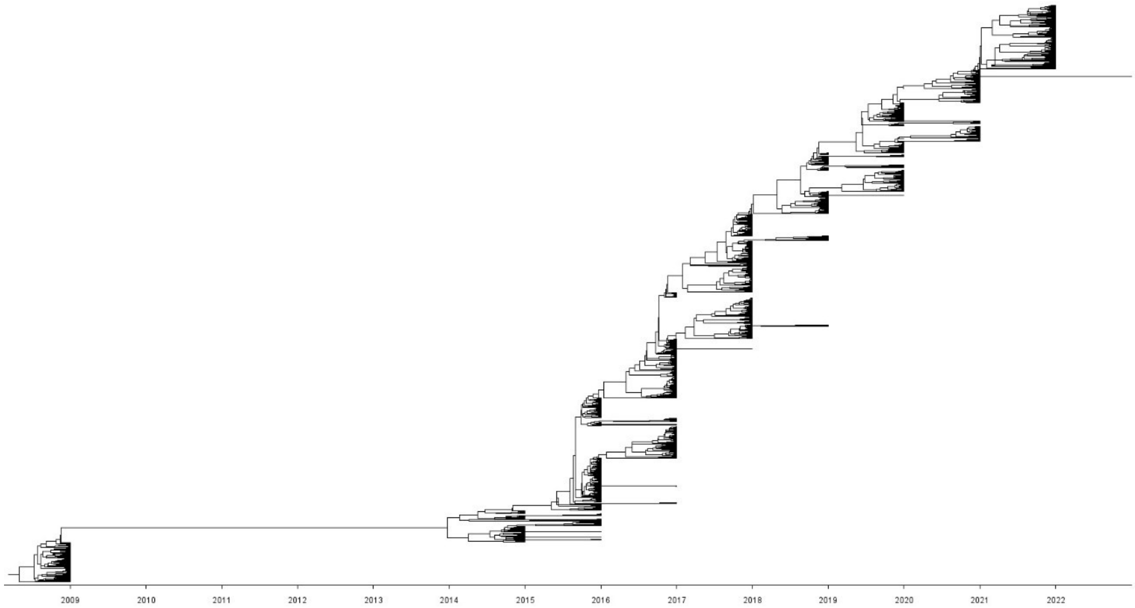


Figure 3.4 The 2009 pandemic H1N1 time tree showing the phylogenetic relationship between NA nucleotide sequences from Africa continent across nine years (2009 – 2022).

The topology of the tree is similar to the topology of the main HA tree. While the distinct separation of 2017 cannot be seen, the separation can be observed in the 2018 sequences. The 2022 sequences are not as separated as Africa HA tree.

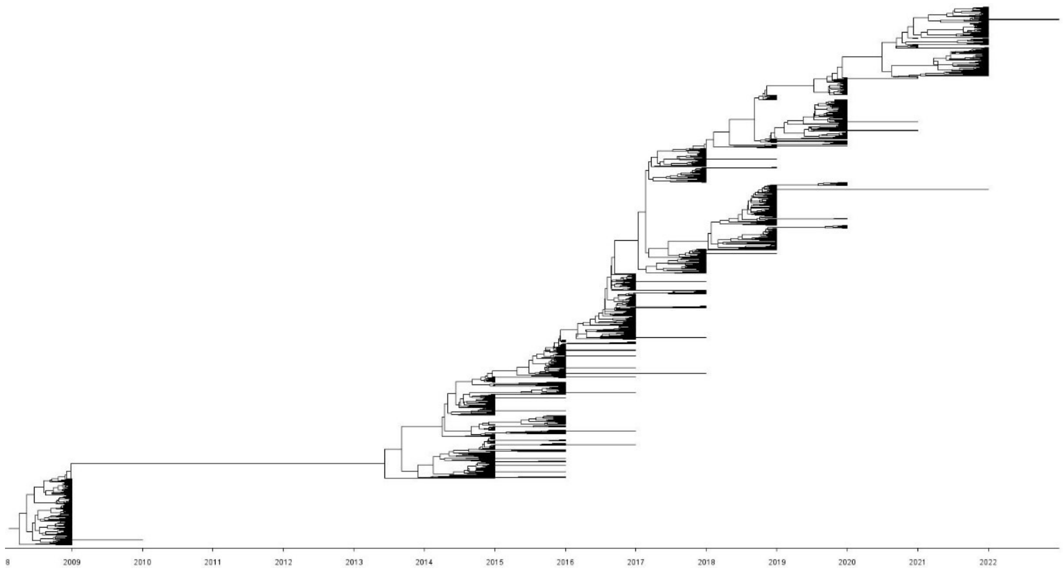


Figure 3.5 The 2009 pandemic H1N1 time tree showing the phylogenetic relationship between HA nucleotide sequences from Asia continent across nine years (2009 – 2022).

The Asia tree for NA was constructed 1610 sequences. Its topology is of similar to the consensus NA tree. Some of the 2021 sequences have their closest relative in 2019, which are similar to the consensus NA tree.

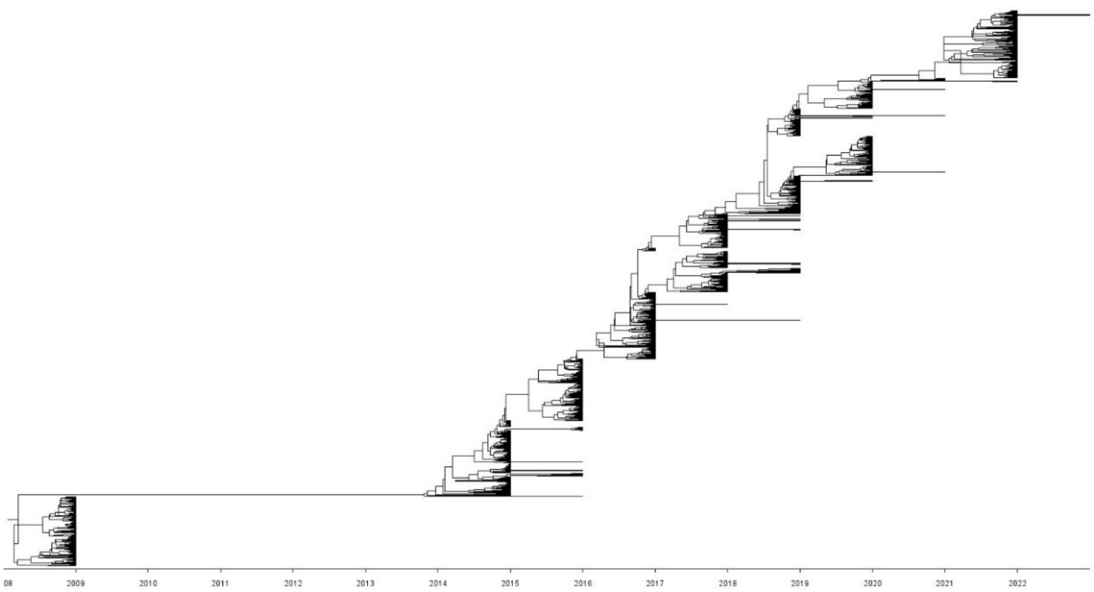


Figure 3.6 The 2009 pandemic H1N1 time tree showing the phylogenetic relationship between NA nucleotide sequences from Asia continent across nine years (2009 – 2022).

The Europe tree topology for HA shows similarity to the consensus HA tree topology. The distinction of two clades in 2017 is observed. The more recent years having longer branch lengths is observed.

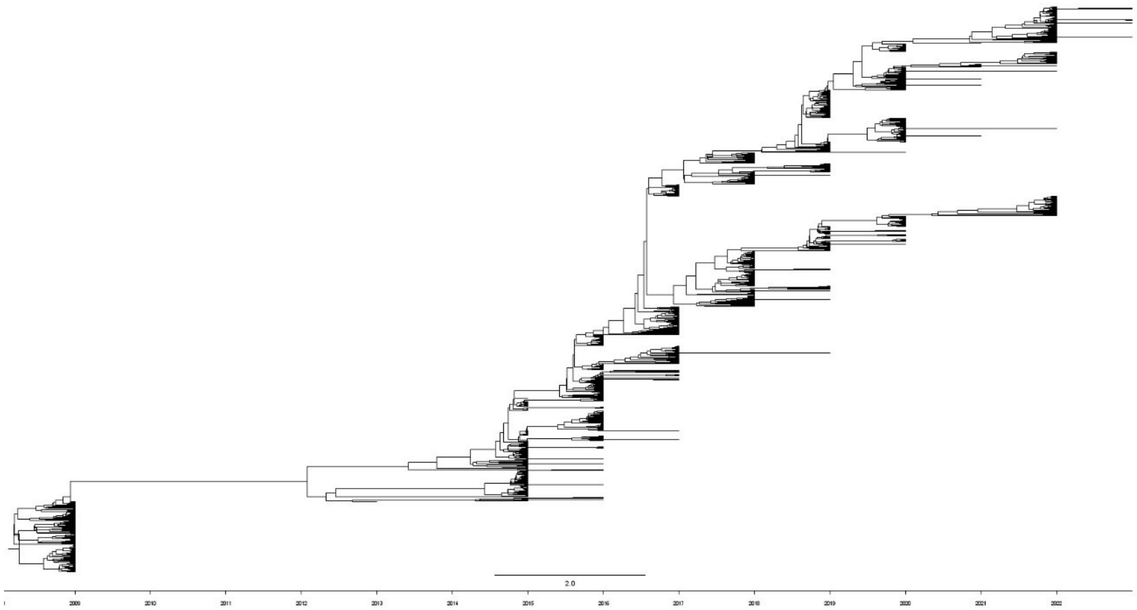


Figure 3.7 The 2009 pandemic H1N1 time tree showing the phylogenetic relationship between HA nucleotide sequences from Europe continent across nine years (2009 – 2022).

The Europe tree for NA shows clade separation of the year 2022, which is similar to the consensus NA tree. The earlier years also show similar topologies with the consensus NA tree.

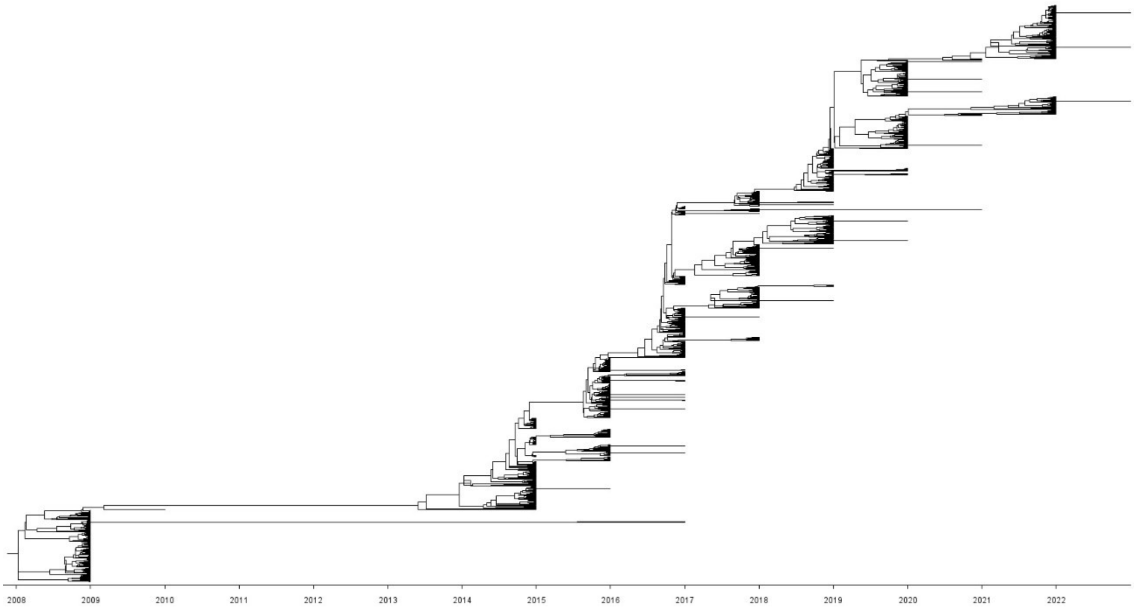


Figure 3.8 The 2009 pandemic H1N1 time tree showing the phylogenetic relationship between NA nucleotide sequences from Europe continent across nine years (2009 – 2022).

The Australia tree for HA shows a more separated topology compared to the consensus HA tree. The smaller 2022 HA clades have a longer branch length than the biggest 2022 clade.

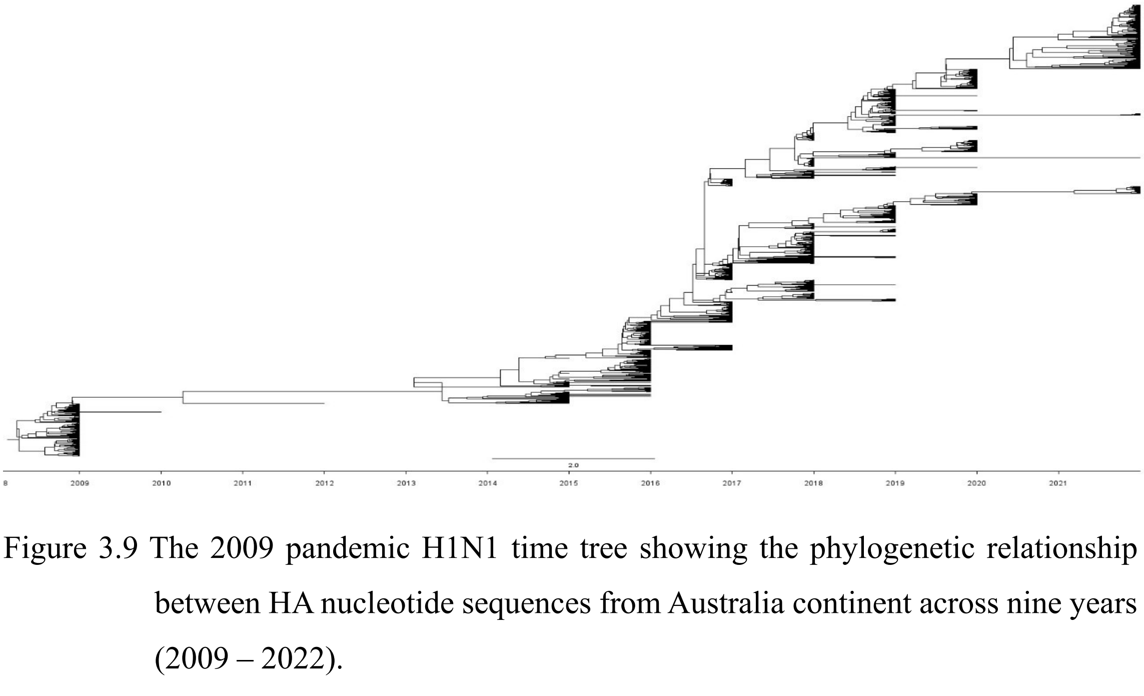


Figure 3.9 The 2009 pandemic H1N1 time tree showing the phylogenetic relationship between HA nucleotide sequences from Australia continent across nine years (2009 – 2022).

All of the 2022 branches have at least a length of two years because there are no available data for 2021 sequences for Australia continent. This is also apparent for the Australia tree for NA. The 2022 sequences are in the same clade except a few, which looks similar to the consensus NA tree.

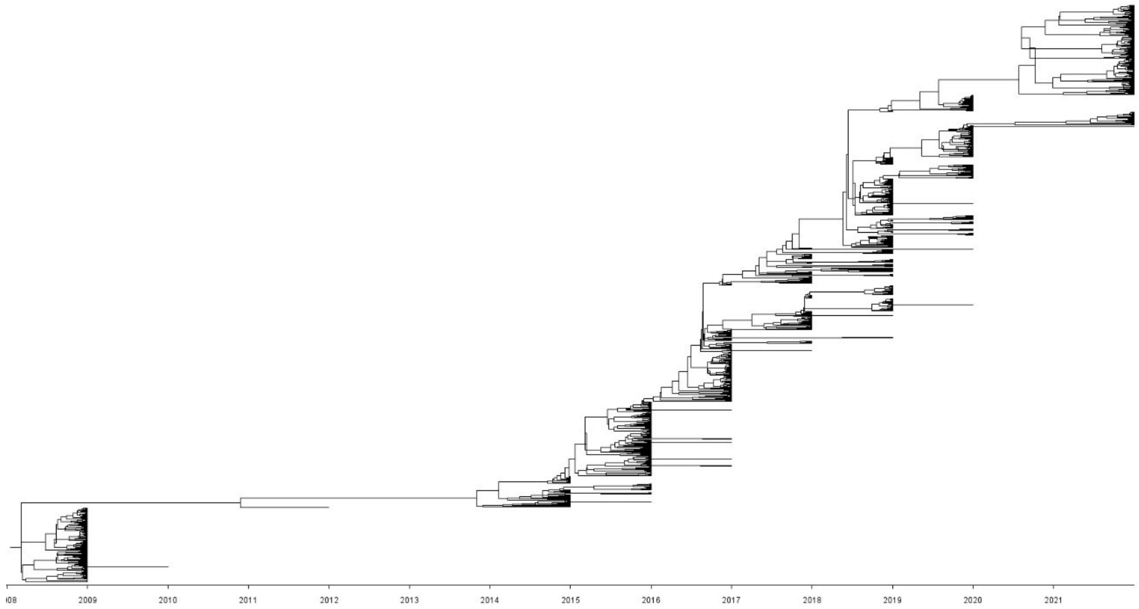


Figure 3.10 The 2009 pandemic H1N1 time tree showing the phylogenetic relationship between NA nucleotide sequences from Australia continent across nine years (2009 – 2022).

The North American tree for the HA nucleotide sequences exhibits a topology that closely resembles the consensus HA tree. Notably, the distinct separation of the clades from the year 2017 is readily discernible. It is noteworthy that one of the larger clades originating in 2022 traces its most recent ancestor back to the year 2020, rather than 2021, a pattern that aligns with the overall consensus HA tree. This characteristic is also observed in the other HA trees, although to a lesser extent when compared to the North American HA tree.

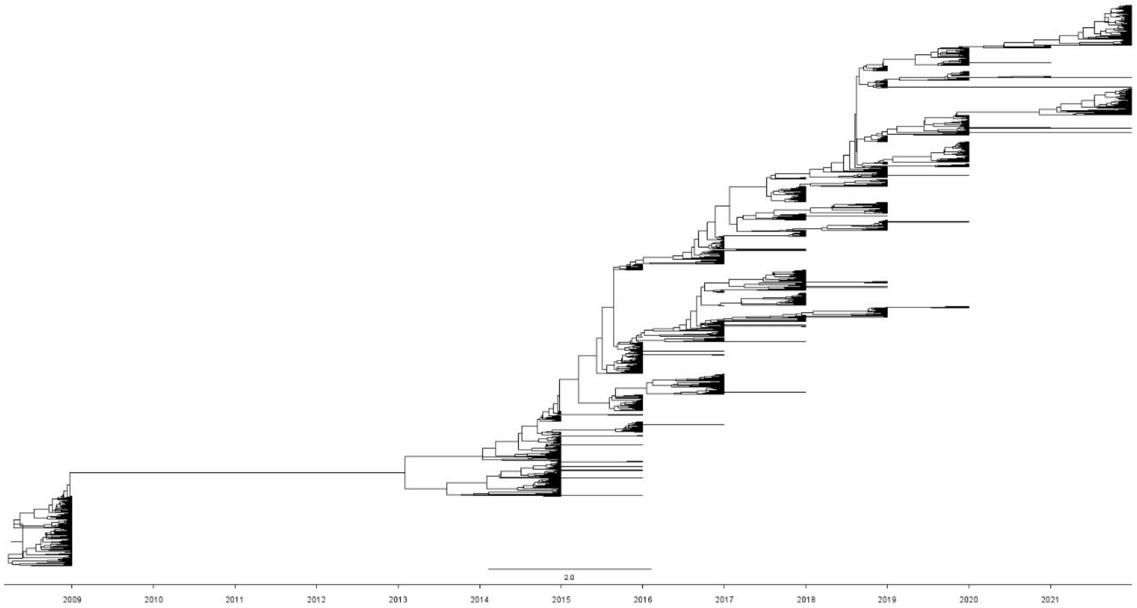


Figure 3.11 The 2009 pandemic H1N1 time tree showing the phylogenetic relationship between HA nucleotide sequences from North America continent across nine years (2009 – 2022).

North American NA tree show a similar topology to the consensus NA tree, the separation of the 2017 clades is similar, and it can be seen that the 2022 sequences cluster together unlike the consensus NA tree.

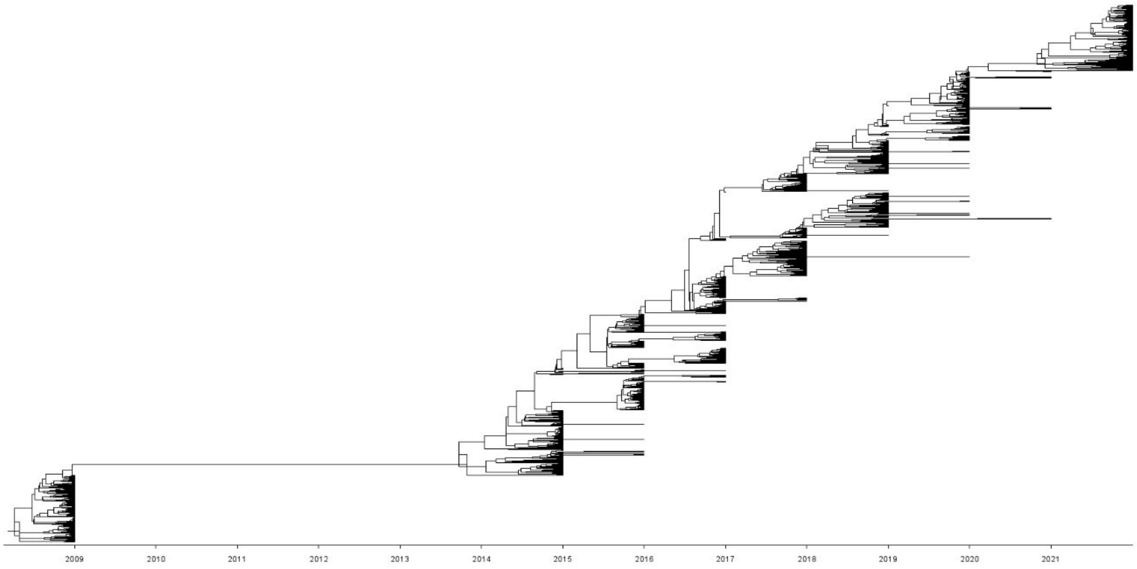


Figure 3.12 The 2009 pandemic H1N1 time tree showing the phylogenetic relationship between NA nucleotide sequences from North America continent across nine years (2009 – 2022).

The South America HA tree topology does not look similar to the consensus HA tree. This might be a result of the nucleotide diversity and selection acting on South America sequences being different from the rest of the regions (Table 3, 4).

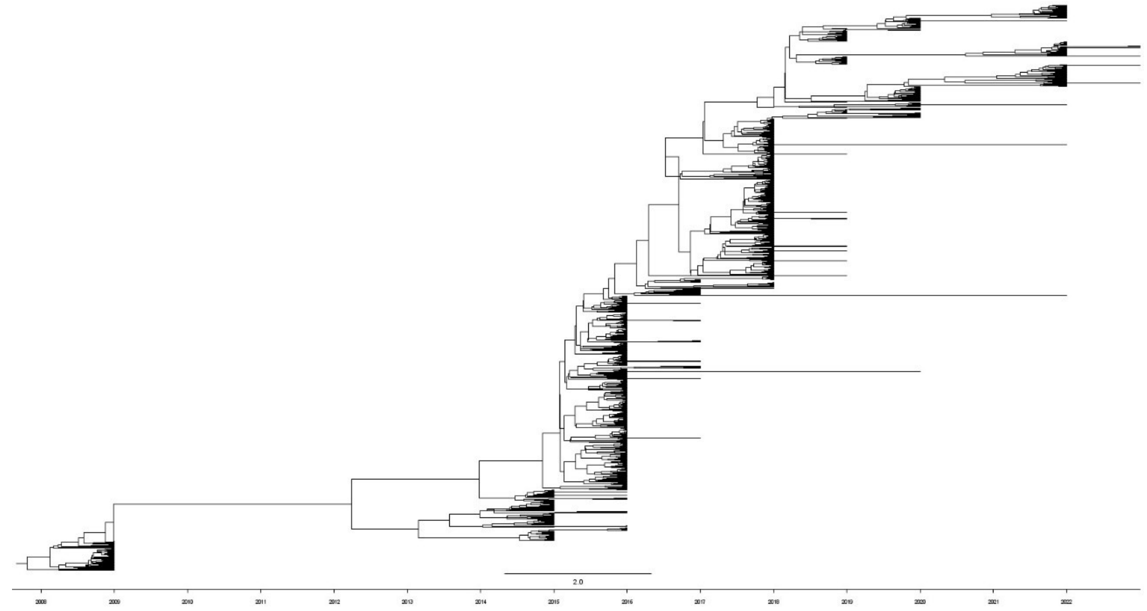


Figure 3.13 The 2009 pandemic H1N1 time tree showing the phylogenetic relationship between HA nucleotide sequences from South America continent across nine years (2009 – 2022).

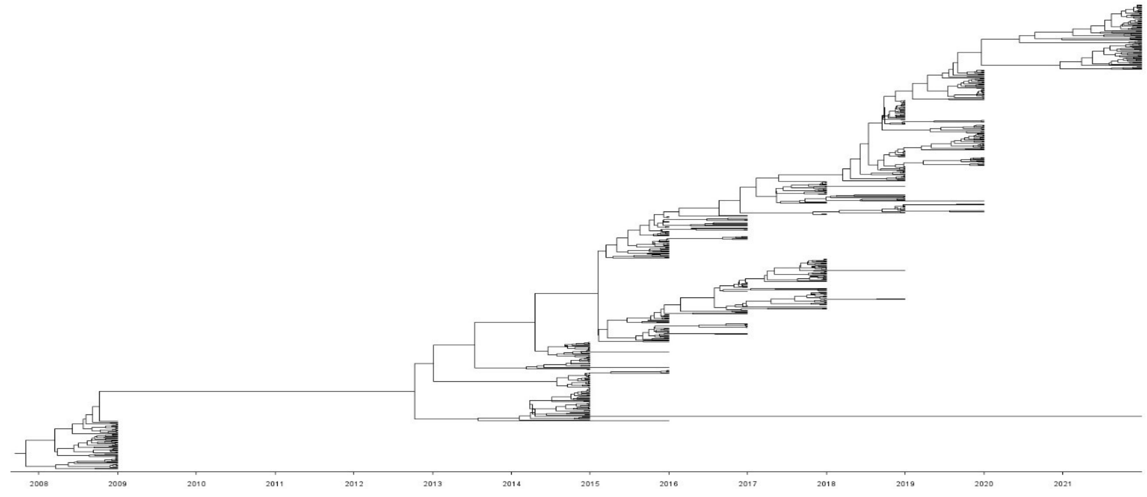


Figure 3.14 The 2009 pandemic H1N1 time tree showing the phylogenetic relationship between NA nucleotide sequences from South America continent across nine years (2009 – 2022).

3.2 Population Genetics and Selection Tests on 2009pdm H1N1 HA and NA Sequences

3.2.1 Genetic Diversity of Hemagglutinin and Neuraminidase Sequences

Year by year comparison of the JC corrected nucleotide diversity of HA shows that the pandemic year of 2009 has the least amount of overall nucleotide diversity compared to the other years (Wilcoxon test $P < 0.001$). This value is higher in 2015 and follows similar trend for the following years. In 2021, the overall nucleotide diversity shows higher variance (spread) compared to the other years, and the overall nucleotide diversity is higher compared to the non-pandemic years. After the unusual 2021 year, the nucleotide diversity of 2022 sequences increased, and the spread of the overall nucleotide diversity is less varied compared to 2021 overall nucleotide diversity.

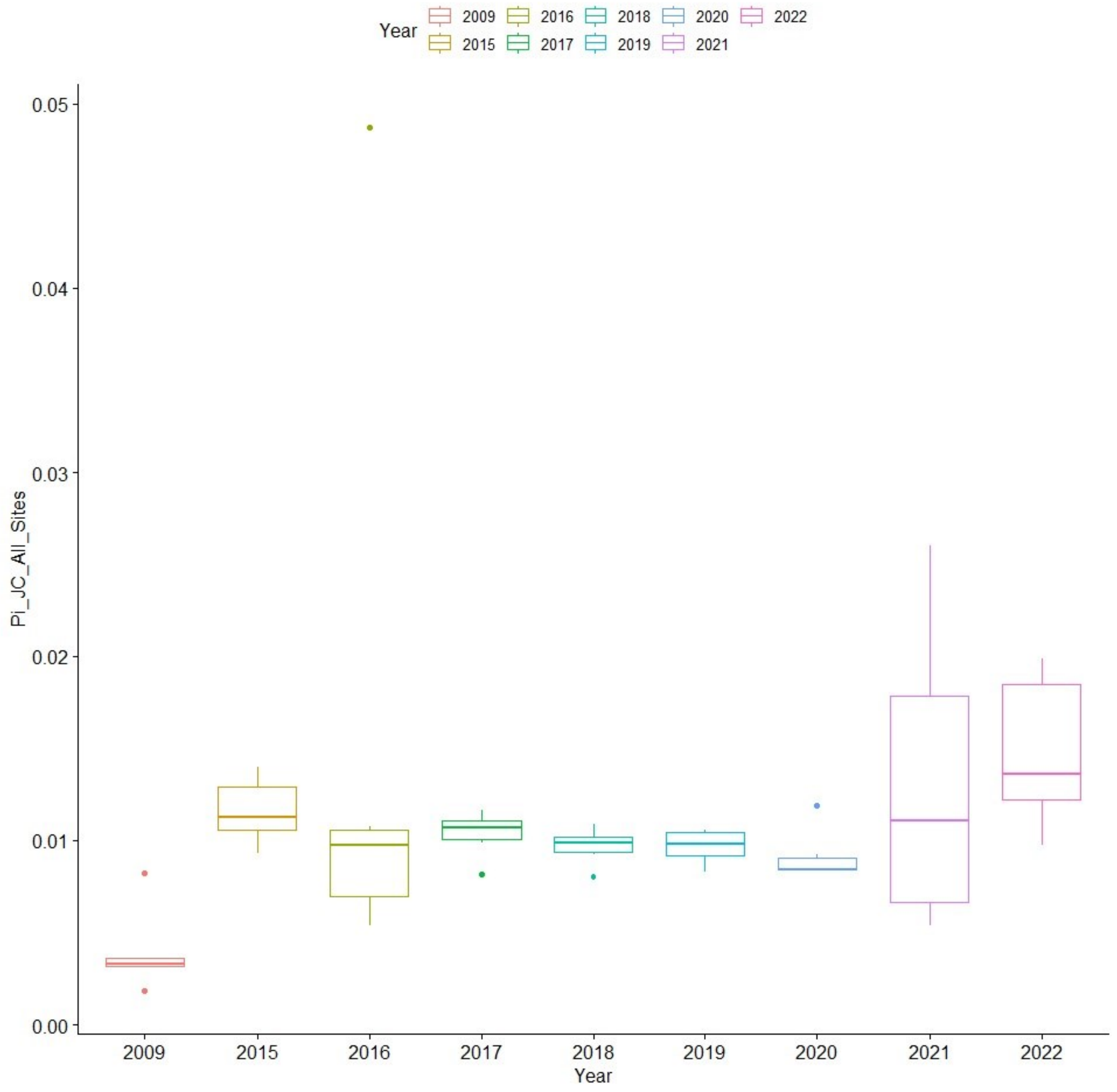


Figure 3.15 The Jukes Cantor corrected nucleotide diversity values of hemagglutinin nucleotide sequences separated by years from 2009 to 2022.

Geographic location wise, the overall nucleotide diversity of all the continents is very similar to each other. There is no statistically significant difference between any of the continents.

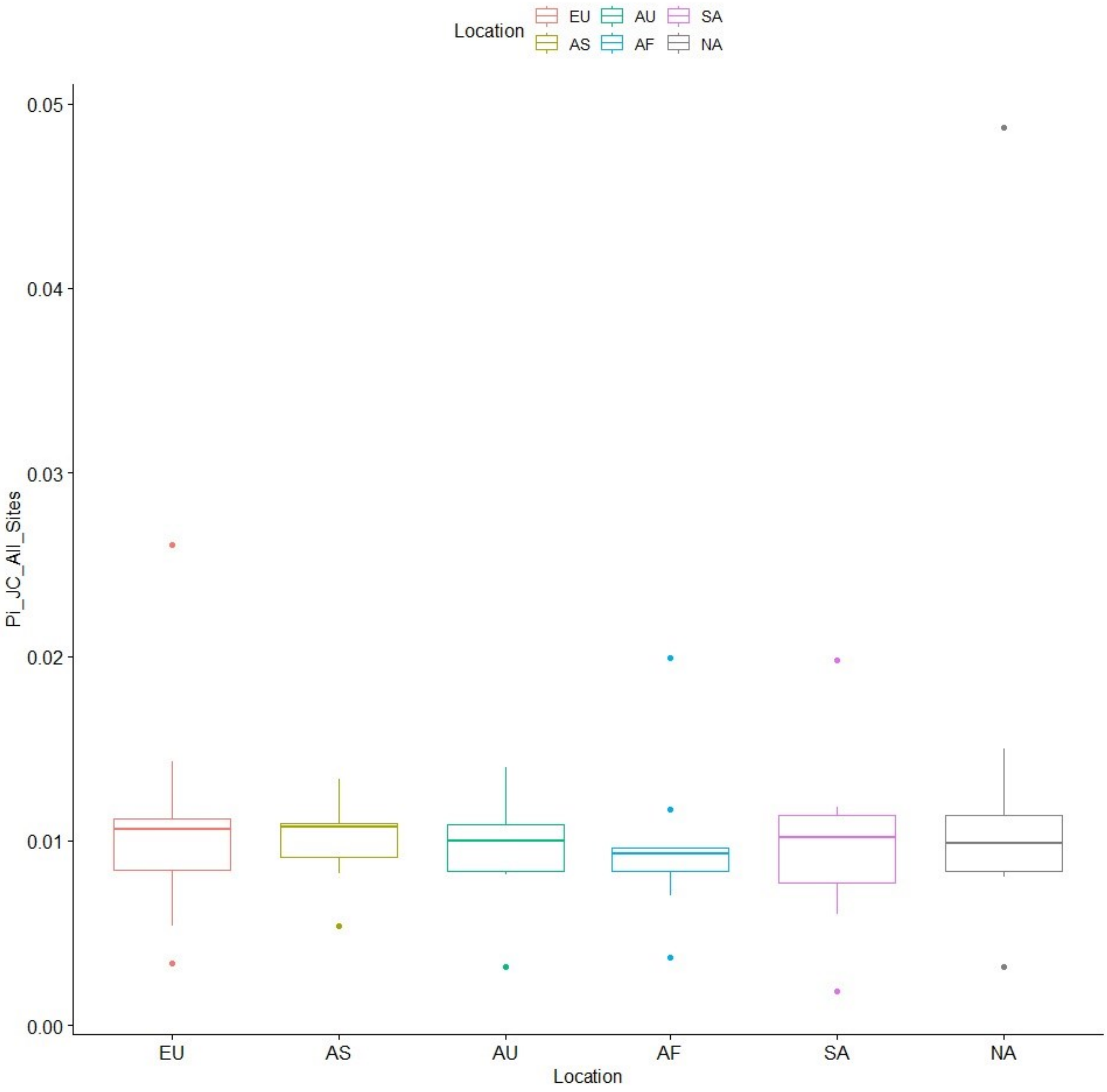


Figure 3.16 The Jukes Cantor corrected nucleotide diversity values of hemagglutinin nucleotide sequences separated by continents.

Looking at the Jukes-Cantor corrected π values, there is no consensus about how the nucleotide diversity is driven. In most of the sequences, the nonsynonymous nucleotide diversity value is higher, however there is a considerable number of sequences where the synonymous nucleotide diversity is higher. For both 2009 and 2022, the HA nucleotide diversity is primarily driven by synonymous diversity rather than nonsynonymous nucleotide diversity (Supplementary Table 1). It is seen that the pandemic year 2009 has the lowest amount of synonymous nucleotide diversity and this value greatly increases after pandemic years. In 2021, the synonymous nucleotide diversity is high but the data variation is very high. In 2022, the variation of the data is less, leading to more statistically significant results.

Specifically, the JC corrected synonymous nucleotide diversity for HA follows a similar trend to the overall nucleotide diversity. The pandemic year of 2009 has the lowest synonymous nucleotide diversity, and the non-pandemic years are higher in value. The synonymous nucleotide diversity being low in 2009 could be attributed to the higher positive selection acting on those sequences. The synonymous nucleotide diversity values start to decrease as the time goes on, from 2015 to 2020, where it is the lowest for non-pandemic years. In 2021, the synonymous nucleotide diversity has increased, and this increase gets more statistically meaningful in 2022.

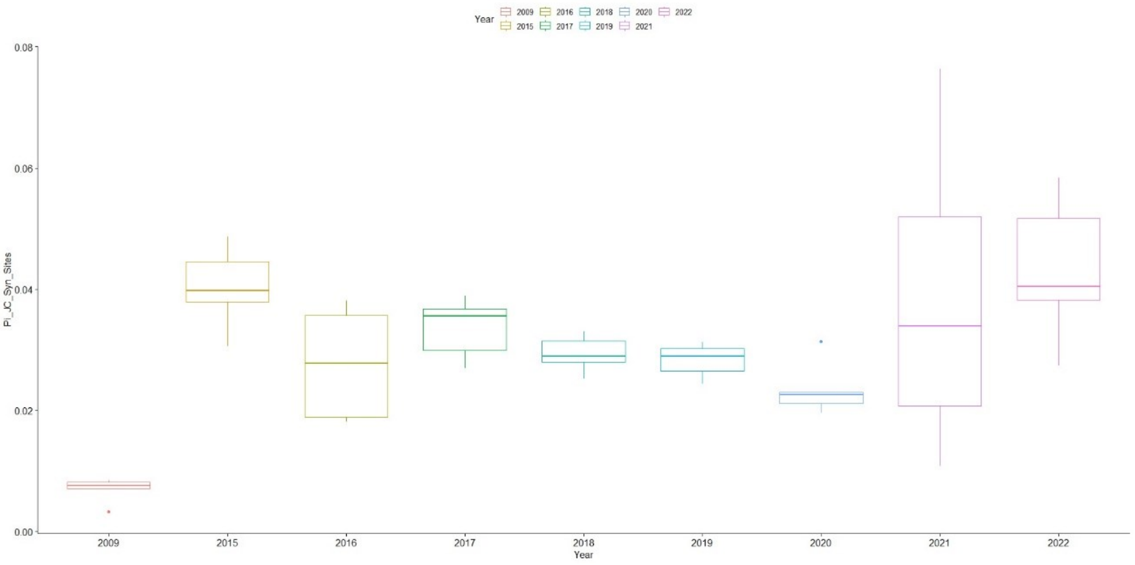


Figure 3.17 The Jukes Cantor corrected synonymous nucleotide diversity values of hemagglutinin nucleotide sequences separated by years from 2009 to 2022.

The continents all have very similar values for the synonymous nucleotide diversity values for the HA protein. There is no statistically significant difference between any of the continents. All of the values for synonymous nucleotide diversity for all the continents are between 0.02 and 0.04. Because synonymous changes do not lead to amino acid changes they do not affect the structure of the molecule, they are expected to be selectively neutral. Therefore, it is not surprising that the 2009 pandemic year sequences show the least amount of synonymous nucleotide diversity. The effect of continent on the synonymous nucleotide diversity cannot be concluded from these findings.

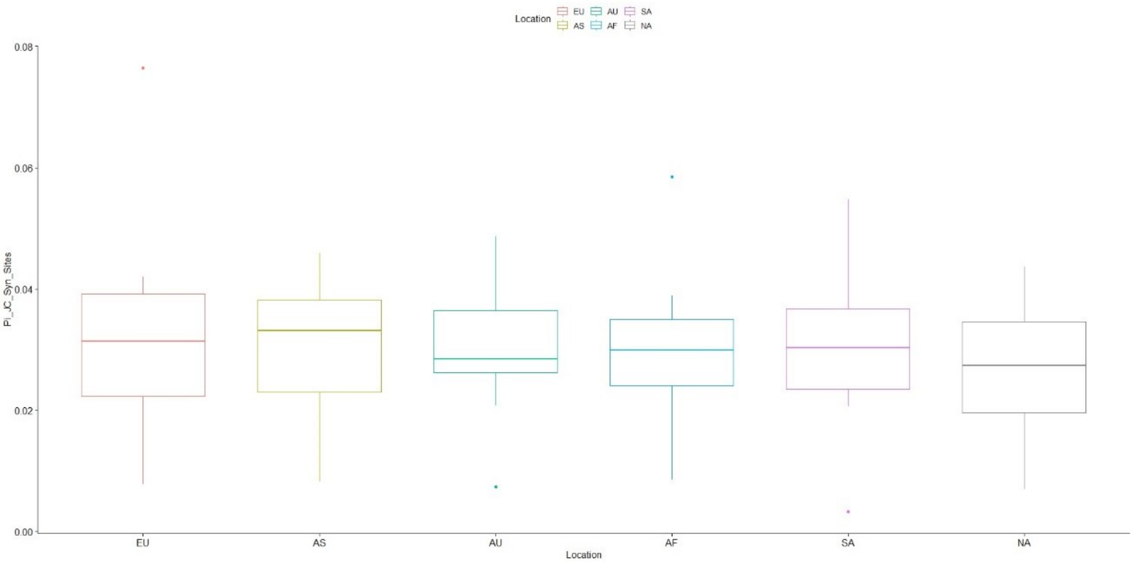


Figure 3.18 The Jukes Cantor corrected synonymous nucleotide diversity values of hemagglutinin nucleotide sequences separated by continents.

The JC corrected nonsynonymous nucleotide diversity for HA shows that the pandemic year of 2009 has the lowest value. Similar to synonymous nucleotide diversity, the values are higher in non-pandemic years with the exception of year 2016. While still higher than 2009, the year 2016 does not follow the previous trend and is lower compared to other non-pandemic years. The nonsynonymous nucleotide diversity keeps increasing from 2015 to 2022, except 2016. In 2021, there is a dramatic increase compared to the previous year. 2022 nonsynonymous nucleotide diversity is also higher than the other non-pandemic years, signalling that visible changes in nucleotide sequences are happening.

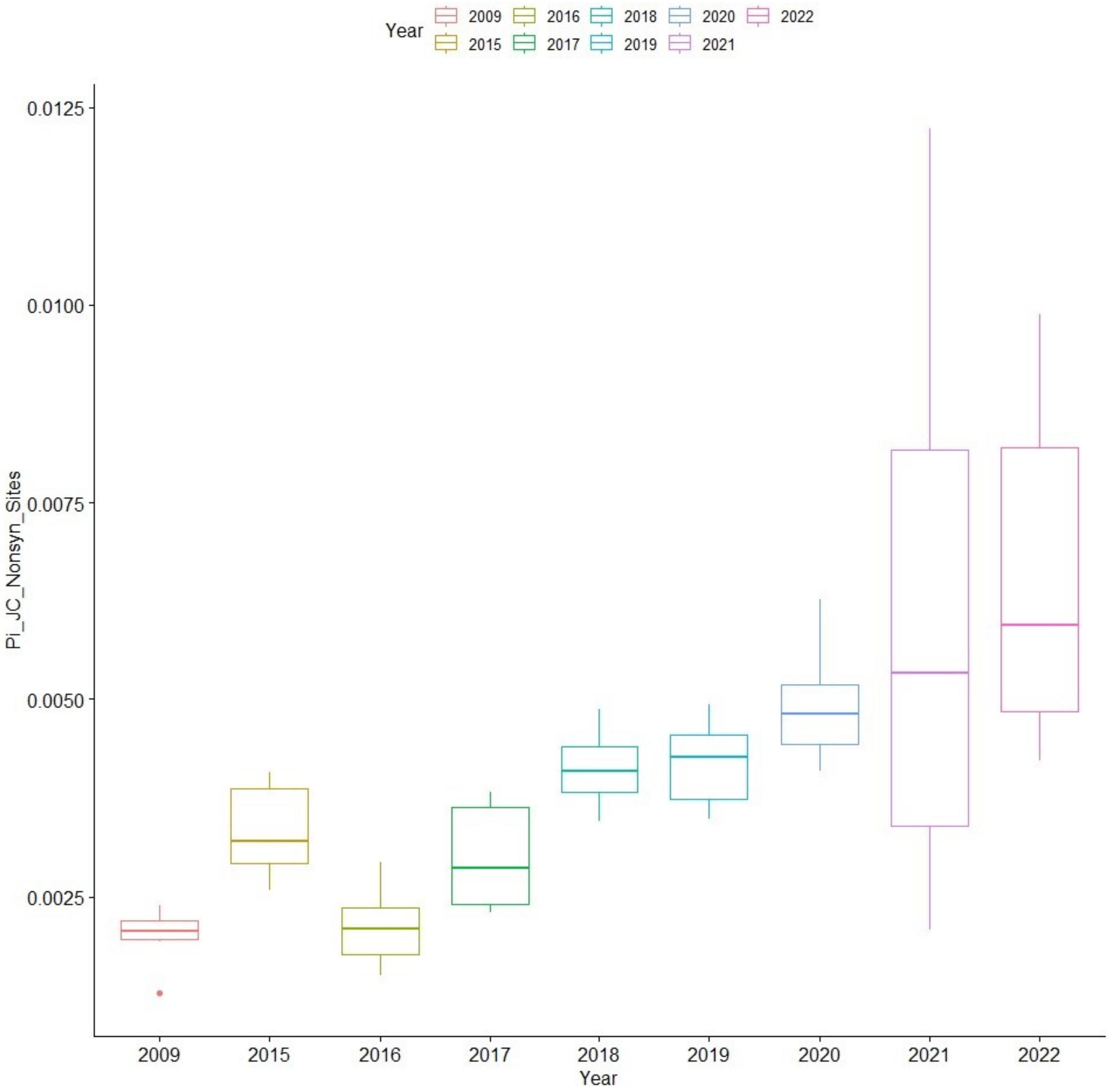


Figure 3.19 Jukes Cantor corrected nonsynonymous nucleotide diversity for hemagglutinin nucleotide sequences separated by years from 2009 to 2022.

The nonsynonymous nucleotide diversity for Africa is lower compared to the other continents. While Africa is statistically significant in its low value, the other continents are not significantly different from each other. This shows that the HA sequences in Africa continent may be not changing as much as the other continents. However the other “small” continents (Australia and South America) are not significantly different from other continents therefore the reason of Africa having lower nonsynonymous nucleotide diversity cannot be explained with lower sample numbers.

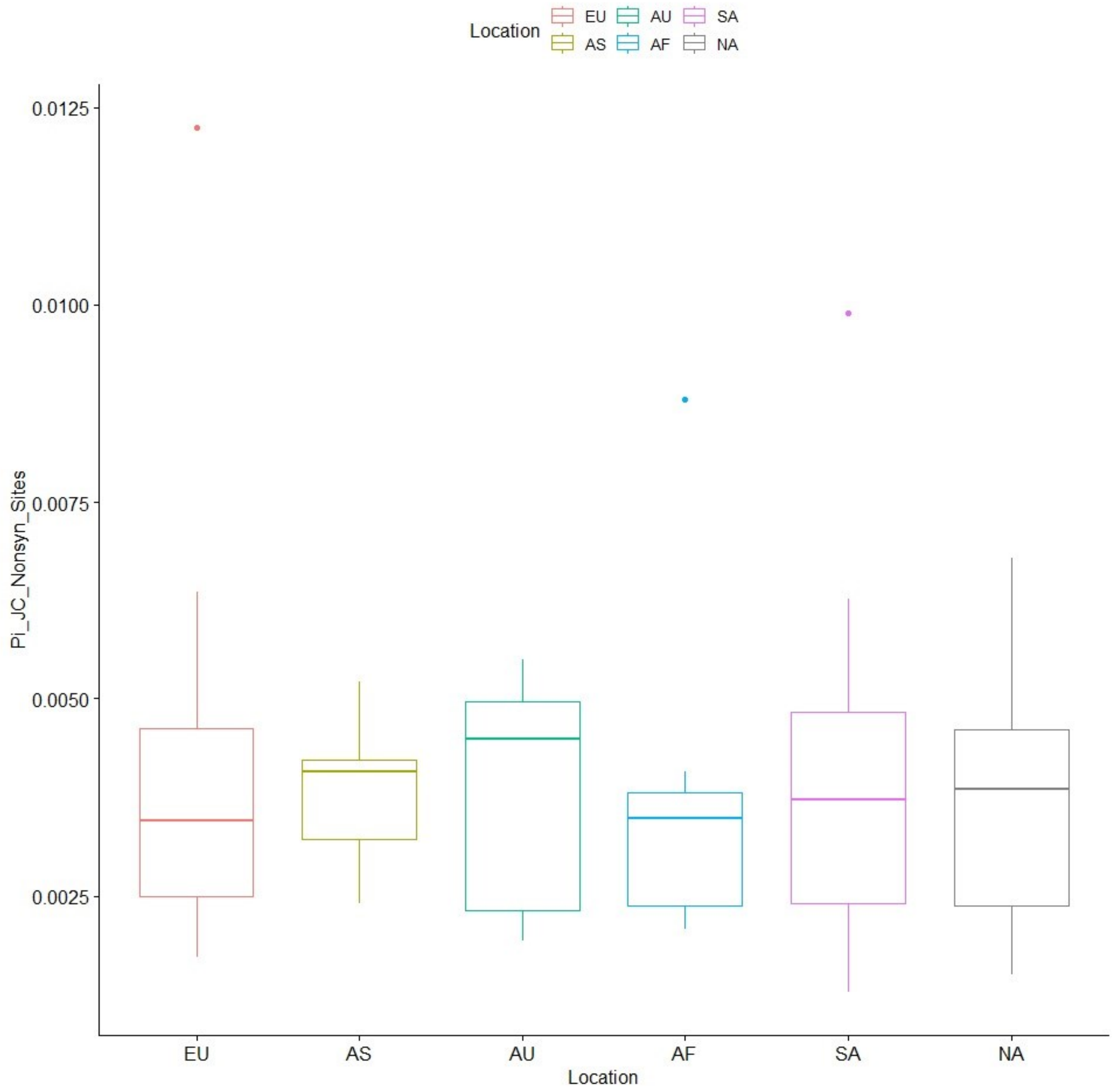


Figure 3.20 Jukes Cantor corrected nonsynonymous nucleotide diversity for hemagglutinin nucleotide sequences separated by continents.

The overall nucleotide diversity of neuraminidase nucleotide sequences follows a similar trend to that of the overall nucleotide diversity of hemagglutinin nucleotide sequences. The pandemic year of 2009 has the lowest amount of overall nucleotide diversity, and this value increases in non-pandemic years. In 2016, the overall nucleotide diversity drops significantly, however the values continues to rise starting at 2017. In 2021, the overall nucleotide diversity is almost double that of 2009. In 2022, the overall nucleotide diversity decreases, however it is still higher than 2009.

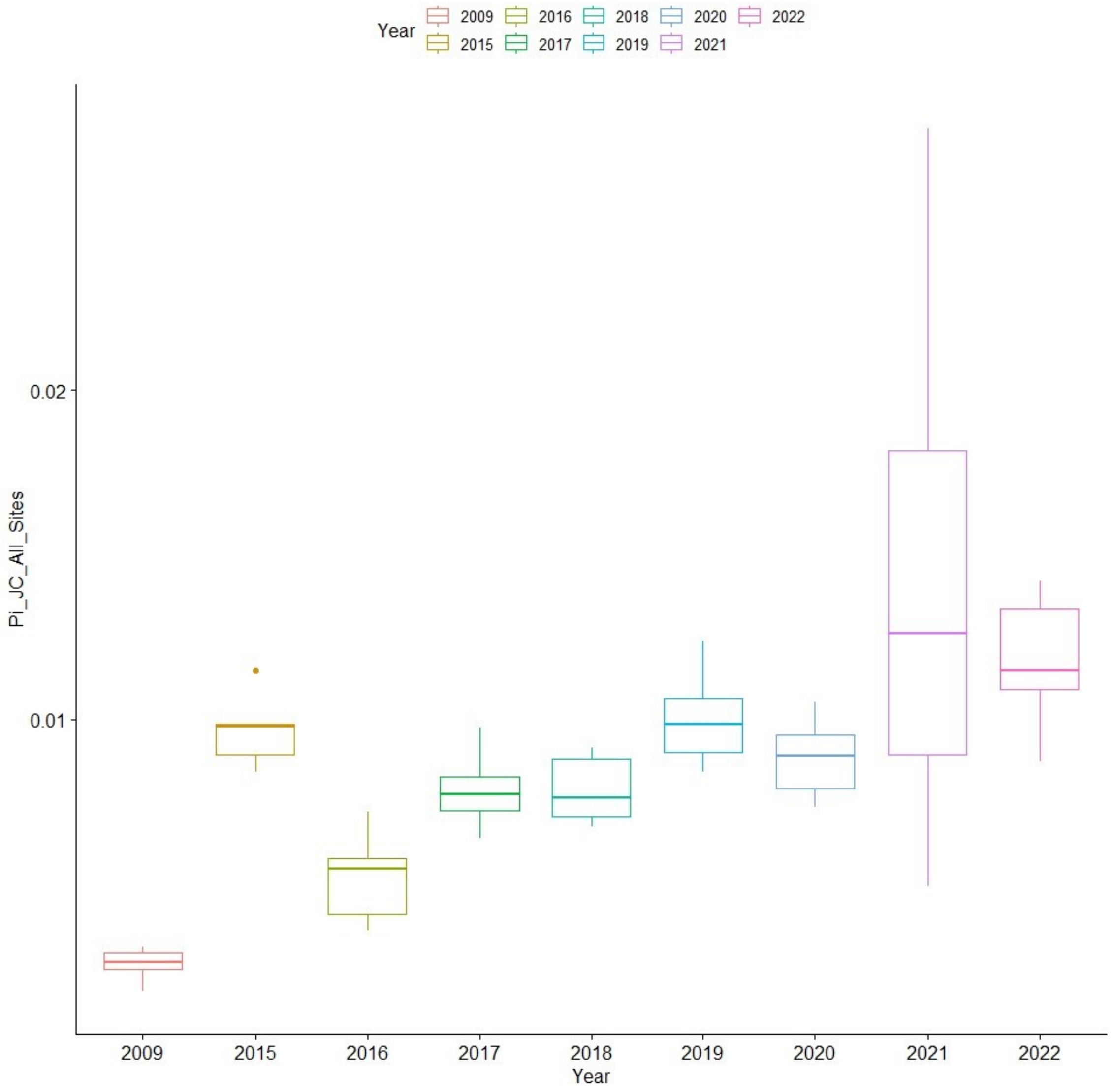


Figure 3.21 The Jukes Cantor corrected overall nucleotide diversity values of neuraminidase nucleotide sequences separated by years from 2009 to 2022.

Overall nucleotide diversity for neuraminidase do not differ significantly continent-wise. Because no continent is different in statistically significant way, which continent has the most overall nucleotide diversity cannot be said.

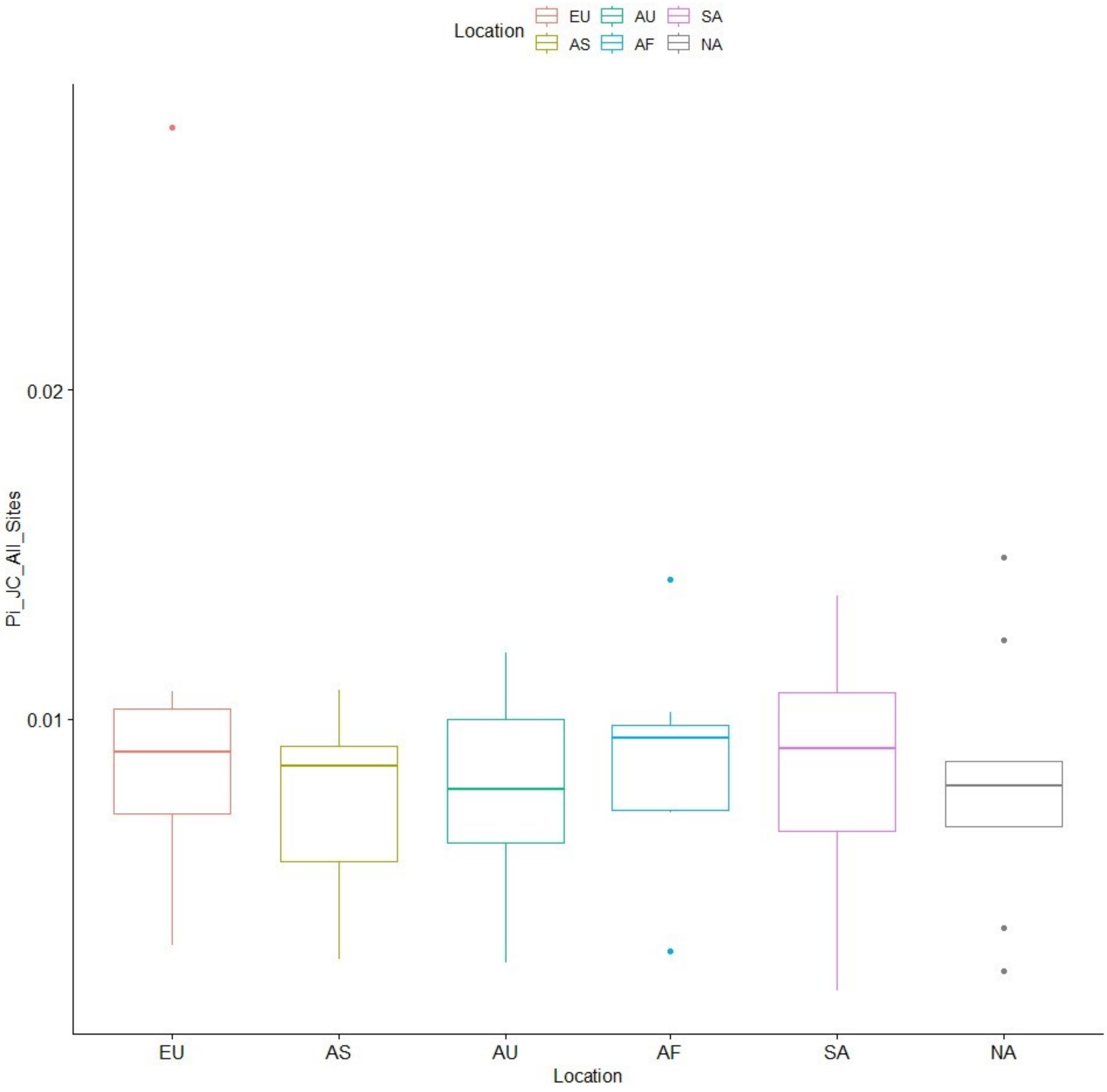


Figure 3.22 The Jukes Cantor corrected overall nucleotide diversity values of neuraminidase nucleotide sequences separated by continents.

The synonymous nucleotide diversity for neuraminidase nucleotide sequences shows that in 2009, the synonymous nucleotide diversity is the lowest. This increases after the pandemic year, and in 2021, the synonymous nucleotide diversity peaks. In 2022, the synonymous nucleotide diversity is still higher compared to 2009 and other non-pandemic years. This suggests that the selection on the neuraminidase molecule is relaxed after the pandemic year, as the synonymous changes are silent changes that do not affect the selection of neuraminidase.

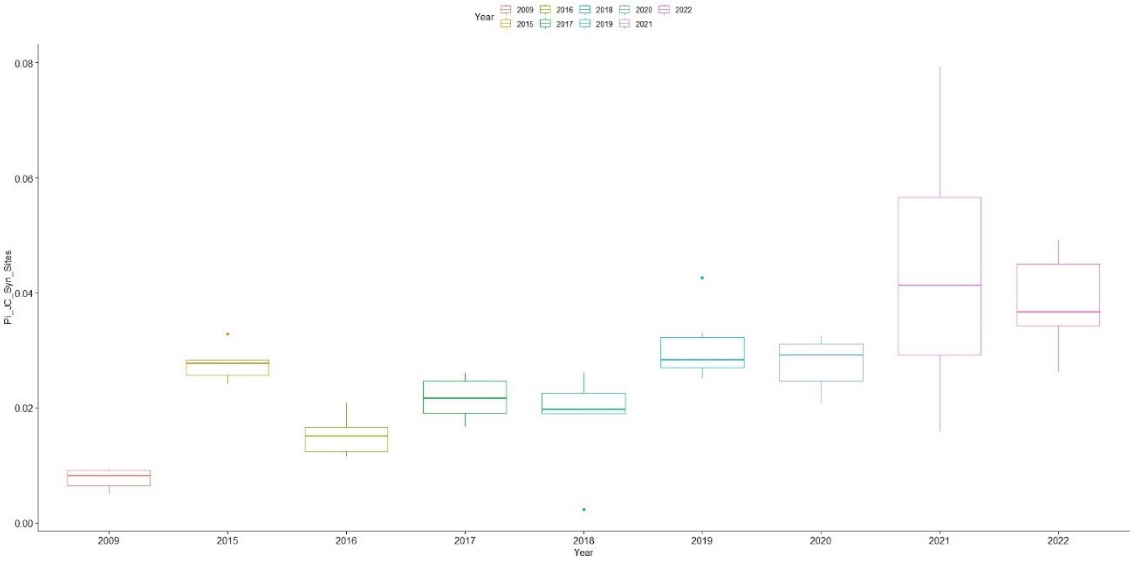


Figure 3.23 The Jukes Cantor corrected synonymous nucleotide diversity values of neuraminidase nucleotide sequences separated by years from 2009 to 2022.

The synonymous nucleotide diversity of neuraminidase does not differ in a statistically significant way between the continents. All the continents are very similar to each other therefore from these values, which continent has the most relaxed selection cannot be inferred.

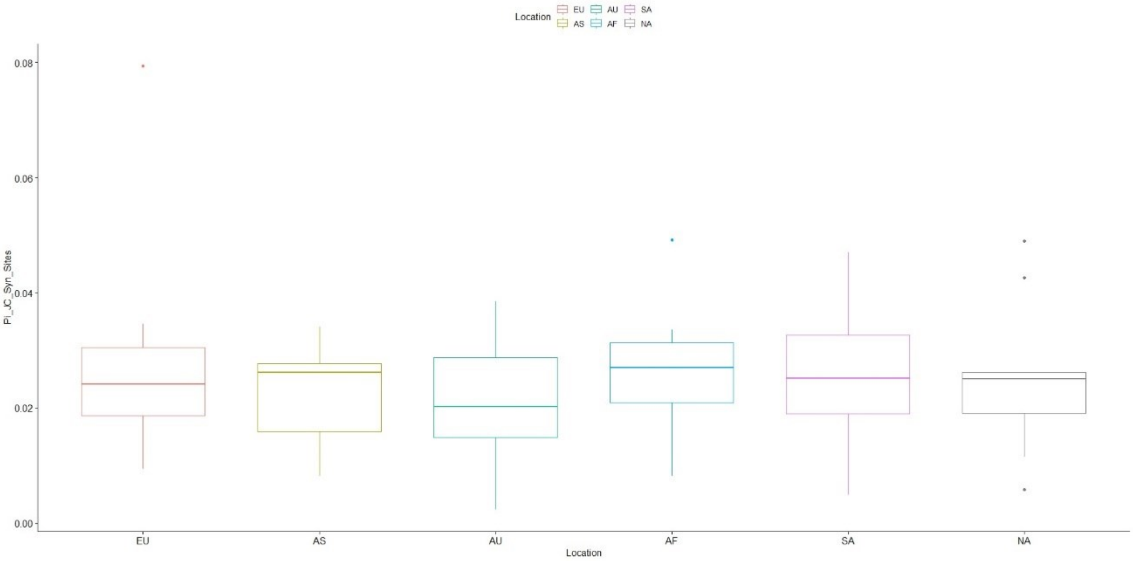


Figure 3.24 The Jukes Cantor corrected synonymous nucleotide diversity values of neuraminidase nucleotide sequences separated by continents.

The pandemic year of 2009 once again shows the least amount of synonymous nucleotide diversity in all of the years. Nonysnonymous nucleotide diversity increases in non-pandemic years. In the year 2021, the nonsynonymous nucleotide diversity is significantly higher than all of the other years; and in 2022 this value lowers back to similar levels of non-pandemic years. While nonsynonymous changes give information about the non-silent changes on the amino acid, the direction of the changes cannot be inferred with this information alone.

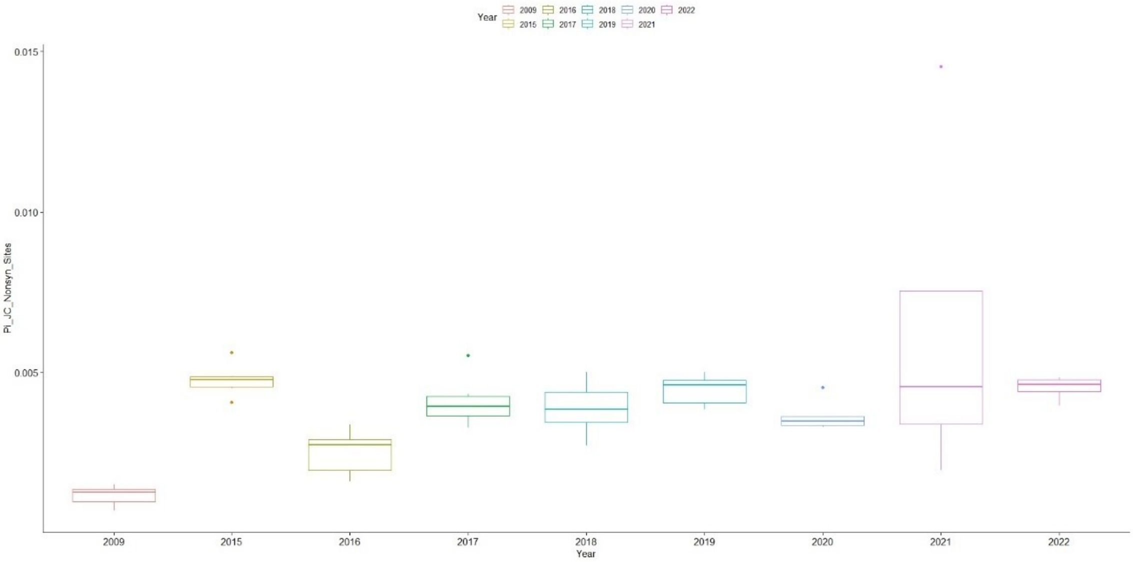


Figure 3.25 The Jukes Cantor corrected nonsynonymous nucleotide diversity values of neuraminidase nucleotide sequences separated by years from 2009 to 2022.

The nonsynonymous nucleotide diversity is very similar when compared continent by continent. The values are not significantly different from each other. Looking at this data, which continent has the amino acid changes that is driven by the nonsynonymous nucleotide diversity cannot be said.

Looking at both HA and NA sequences, it can be seen that the HA polymorphism numbers are on most cases higher compared to the NA samples. Looking at 2009 and 2022 numbers, Europe in 2022 has the highest amount of synonymous polymorphism changes and North America in 2009 has the highest amount of replacement changes. It can be seen that the more populated continents have a higher amount of both synonymous and replacement changes compared to those with less population.

2021 Europe samples for NA have the highest value for nucleotide diversity tests, as well as Jukes-Cantor corrected tests for all kinds of nucleotide diversity. The lowest value of JC corrected nucleotide diversity belongs to the South America in 2009. Comparing the 2009 pandemic year and 2022 post COVID-19 pandemic data, it is seen that 2022 in general has a higher amount of nucleotide diversity in NA sequences. Overall, JC corrected nucleotide diversity values are varying for nucleotide diversity of synonymous sites and nonsynonymous sites, therefore it cannot be decided which kind of nucleotide diversity type is driving the evolution of NA of 2009 pandemic H1N1.

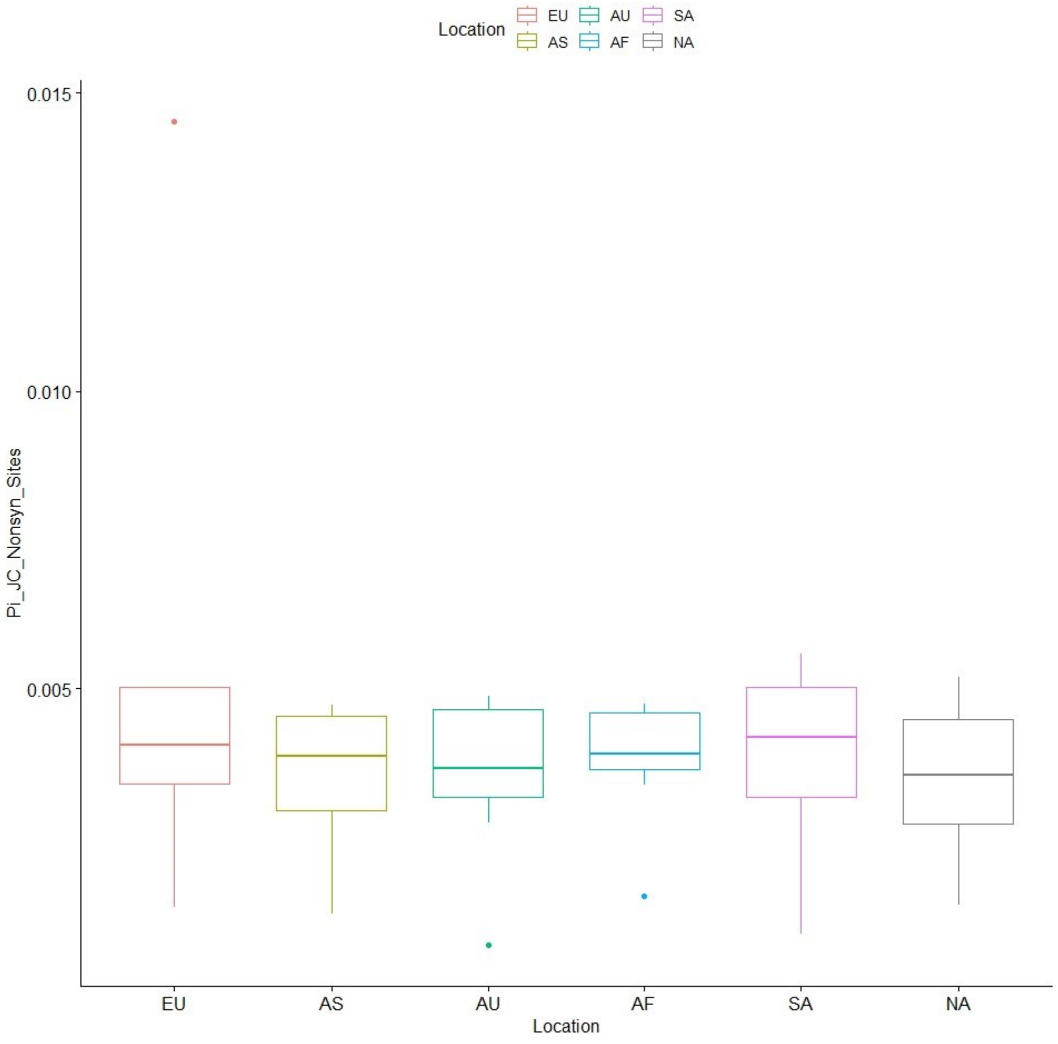


Figure 3.26 The Jukes Cantor corrected nonsynonymous nucleotide diversity values of neuraminidase nucleotide sequences separated by continents.

Looking at the nucleotide diversity for all sites, the highest amount of nucleotide diversity belongs to the 2020 North America sequences, while the least amount belongs to the 2009 South America sequences. The 2022 sequences have higher nucleotide diversity compared to 2009 sequences. In 2009, Asia has the highest π meanwhile South America has the lowest. In 2022, Africa has the highest π meanwhile North America has the lowest value (Supplementary Table 1).

The population genetics tests for NA show that the highest amount of segregating sites belongs to 2022 Europe samples meanwhile the lowest amount belongs to the 2021 Asia continent (Supplemental Table 2). Between the 2009 and 2022 NA sequences, there is no clear distinction as to which group has the higher segregating sites. Looking at all the years and continents, singleton values are almost always lower than the parsimony informative sites. Comparing the 2009 and 2022 sequences, the 2022 sequences have more singleton and parsimony informative sites, however North America and Asia have more singleton sites in 2009 compared to 2022. Following the trend of HA, the populations where there are fewer samples such as South America, Australia, and Africa of some years, the singleton sites are greater than the parsimony informative sites. For all of the NA sequences, the synonymous polymorphisms are greater than the replacement polymorphisms (Supplemental Table 2).

3.2.2 Selection Tests on Hemagglutinin and Neuraminidase Sequences

Performing the neutrality tests for HA and NA nucleotide sequences helps with the assessment of selection that is acting on the molecules. All of the Tajima's D tests are negative with almost all of them having statistical significance (Supplementary Table 3, 4). The Fu-Li's D* and F* tests are in line with the results of Tajima's D tests. In all of the years, 2009 had the most negative TD results, meanwhile 2021 had the least negative results and 2021 results are not statistically significant (Supplementary Table 3, 4). Comparing 2009 and 2022, year 2009 has the more negative values for the neutrality tests (Supplementary Table 3, 4).

The negative values of these tests signify that there is a large amount of rare variants in the samples rather than the consensus variants as a result of rapid population growth or a selection on the population.

In 2009, overall TD value of hemagglutinin for pandemic year 2009 is the lowest, which indicates that the selection acting on these hemagglutinin sequences is the most positive and there is a greater population growth in 2009. In non-pandemic years the TD values get closer to zero, which is expected compared to the pandemic years. In both 2015 and 2021, the TD values are higher than the other non-pandemic years which signals the selection is relaxed and the population might be not growing as much compared to other pandemic and non-pandemic years. After 2021, the TD values are going down, meaning the positive selection on the hemagglutinin sequences is returning, and the population of influenza may be growing compared to 2021.

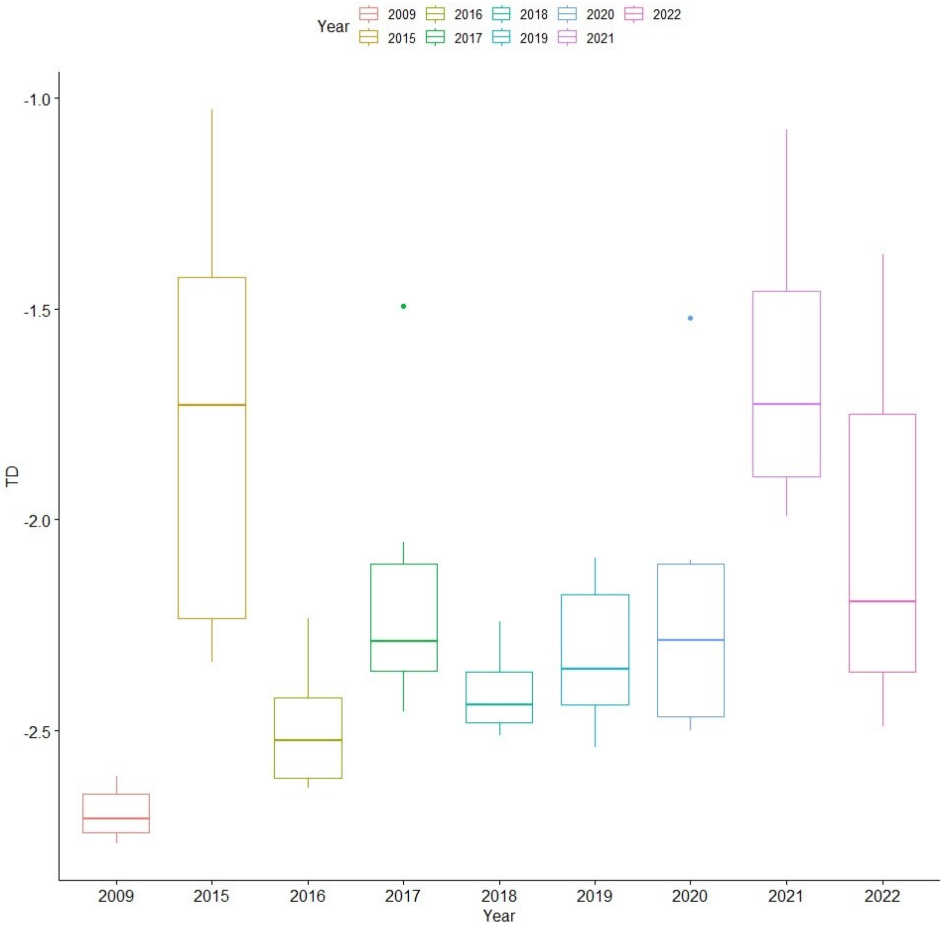


Figure 3.27 The overall Tajima's D test results of hemagglutinin nucleotide sequences separated by year from 2009 to 2022.

Location-wise, the continents with lower amount of samples (Africa, South America, Australia) have higher TD values, which indicates that the selection on hemagglutinin on these continents are more relaxed compared to the big continents (Asia, Europe, North America). Higher TD results also mean that the population growth on these continents may be lower compared to the continents that have more negative TD values. The TD values indicate that in the bigger continents, the virus is effected by more positive selection and its population growth is higher compared to the smaller continents.

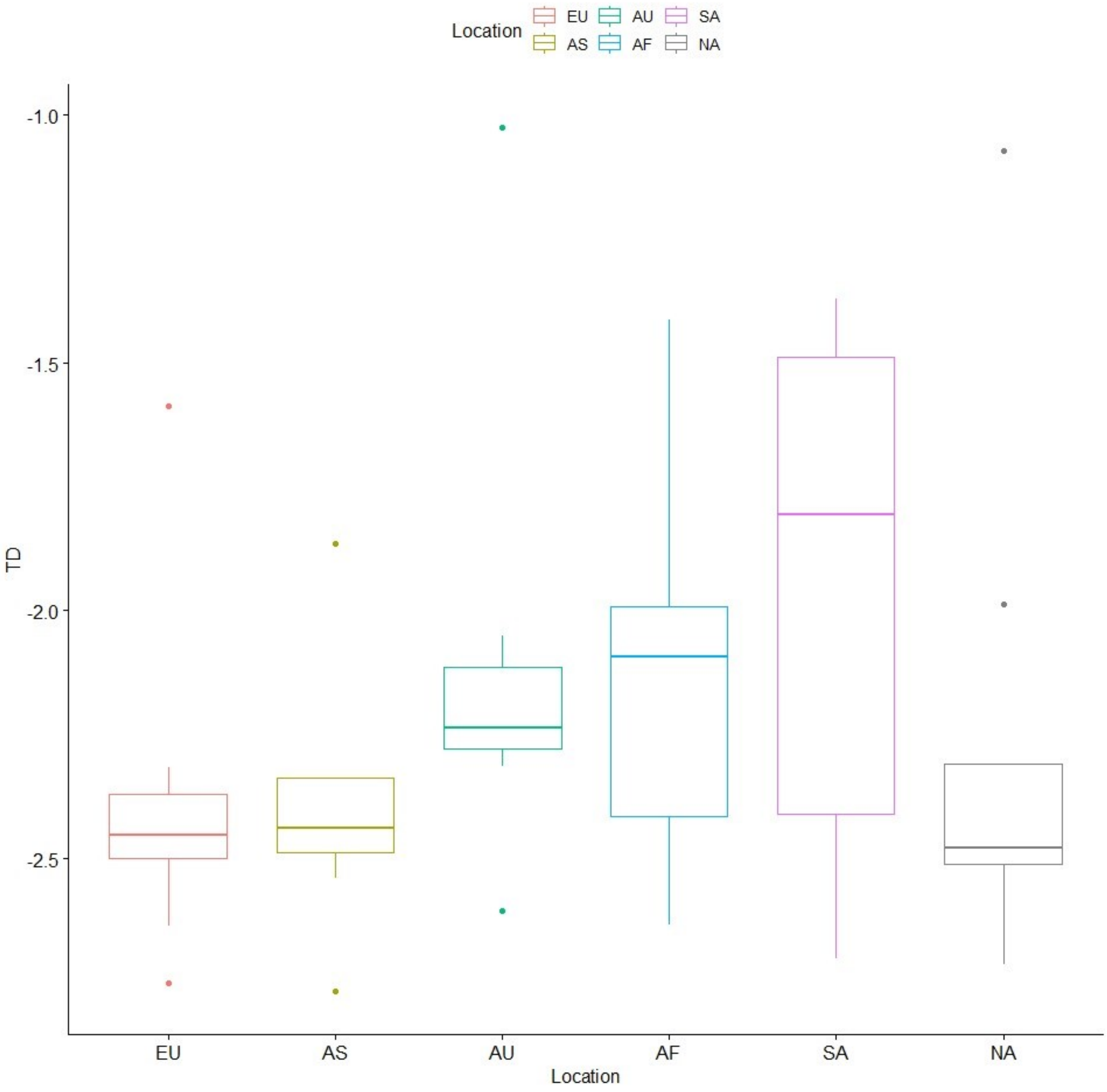


Figure 3.28 The overall Tajima’s D test results of hemagglutinin nucleotide sequences separated by continents.

The nonsynonymous TD gives further information about the selection acting on the sequences. The pandemic year of 2009 has the most negative nonsynonymous TD value, indicating that 2009 has the most positive selection acting on the sequences. This selection gets relaxed for the non-pandemic years in comparison. In 2021, the nonsynonymous TD value is the least negative, indicating that the selection on hemagglutinin is the most relaxed in that year. In 2022, the sequences are getting affected by more positive selection compared to the previous years, which means the changes on the amino acid sequences are beneficial to hemagglutinin, and to influenza virus.

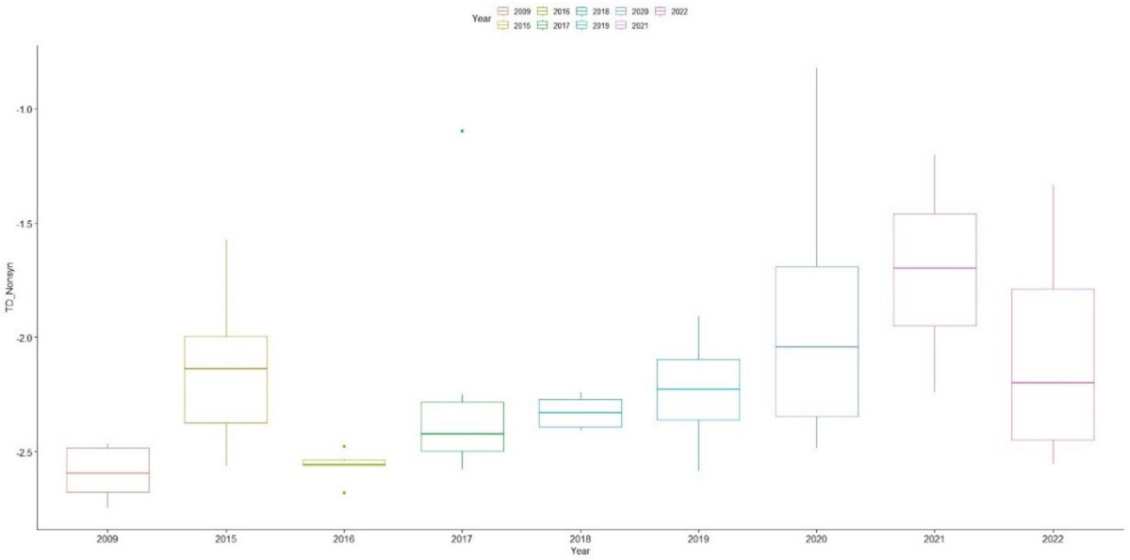


Figure 3.29 The nonsynonymous Tajima’s D test results of hemagglutinin nucleotide sequences separated by years from 2009 to 2022.

The results for the nonsynonymous TD values show that the big continents (Asia, Europe, North America) have significantly more negative TD values than the other three continents (Wilcoxon test $P < 0.05$). This shows that the negative overall TD values are supported by the nonsynonymous TD values, meaning there is positive selection affecting the sequences and negative TD values are not just a product of fast population growth.

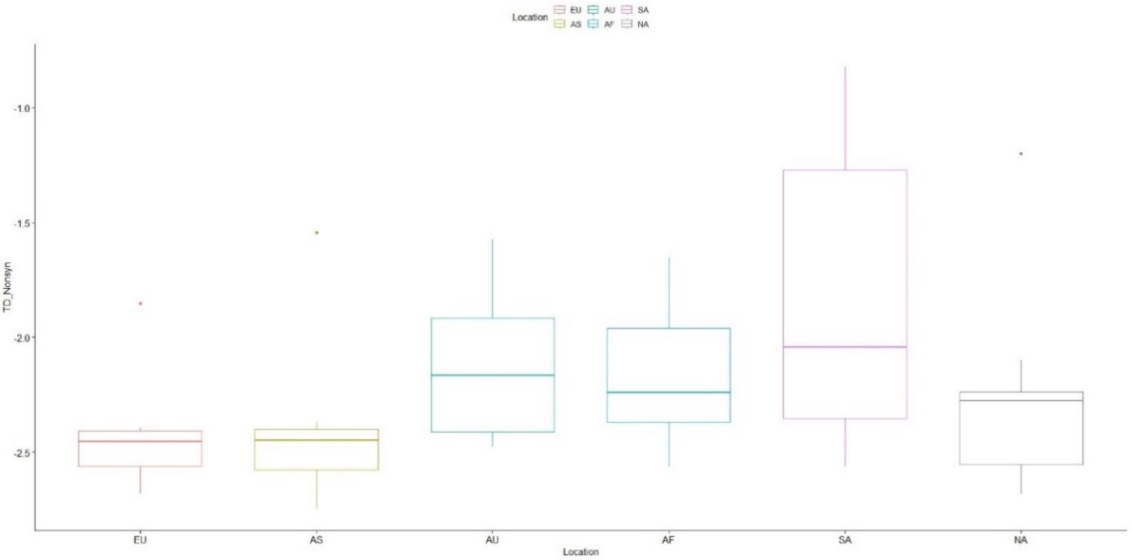


Figure 3.30 The nonsynonymous Tajima’s D test results of hemagglutinin nucleotide sequences separated by continents.

The synonymous Tajima’s D values give information about the selection and population growth on the molecule based on the synonymous changes in the nucleotide sequences. However, the synonymous changes on the nucleotide sequences do not change the amino acid sequence, and the protein structure; therefore, the synonymous changes are expected to be neutral (silent). Because these changes are silent, they do not affect the selection on the nucleotide sequences. Instead, the synonymous TD values give information about the population growth.

The pandemic year of 2009 has the most negative synonymous TD values, which indicates that the population growth on that year is the highest compared to the other years.

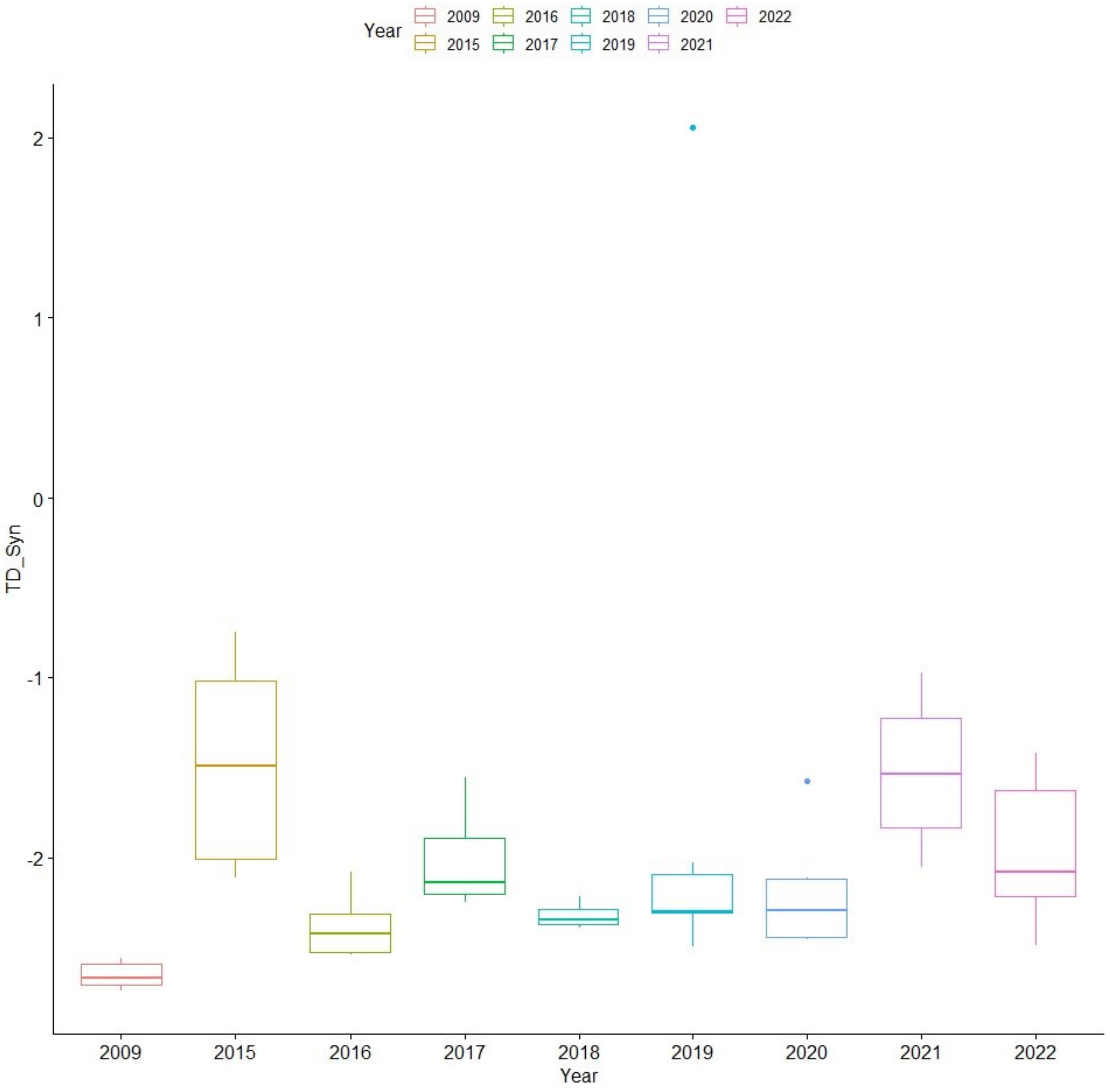


Figure 3.31 The synonymous Tajima's D test results of hemagglutinin nucleotide sequences separated by years continents.

Looking at the synonymous TD values by continent, the big continents (Europe, Asia, North America) have the more negative values, which indicates the growth rate for these continents are higher than the other three.

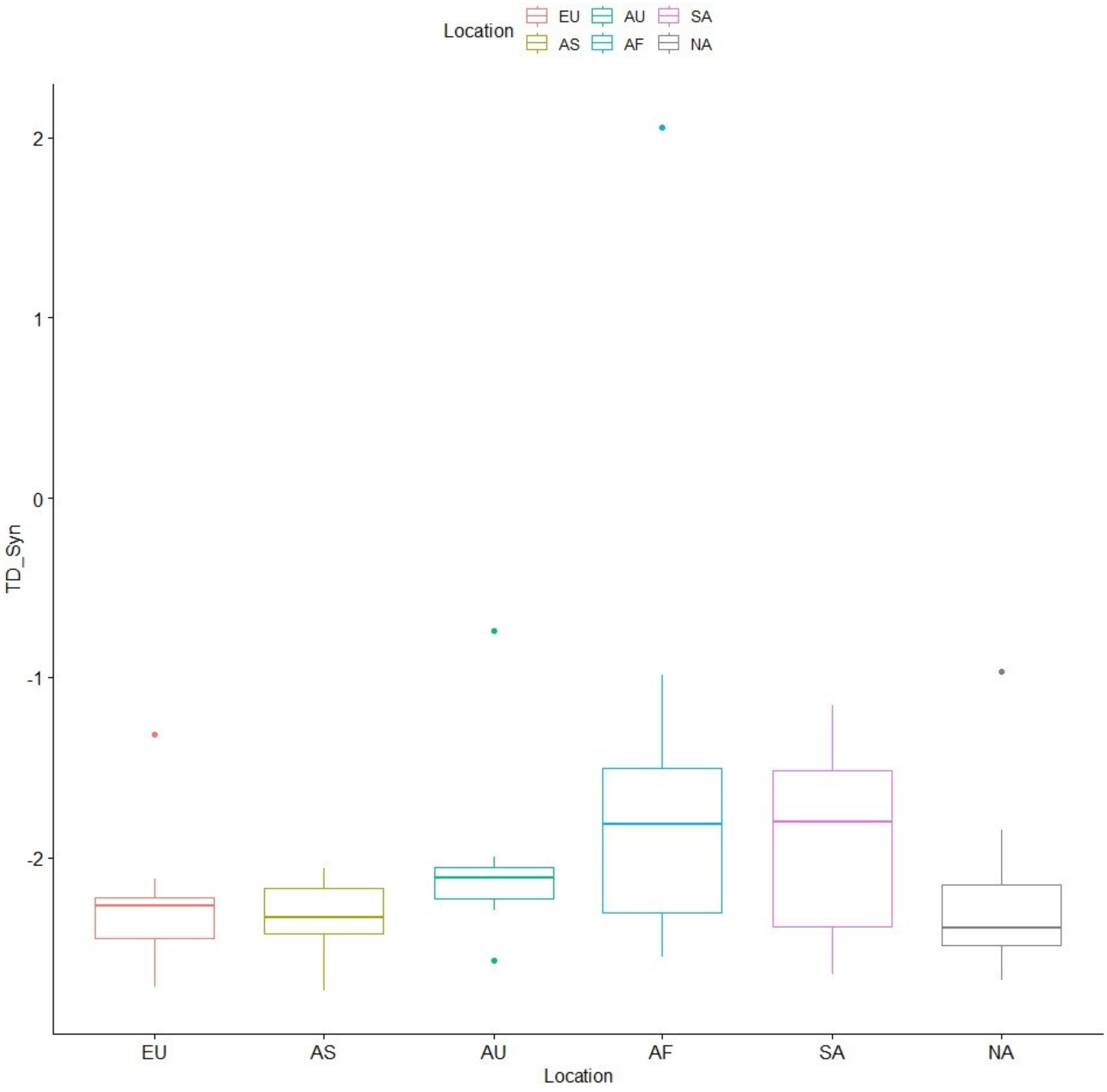


Figure 3.32 The synonymous Tajima’s D test results of hemagglutinin nucleotide sequences separated by years continents.

The Fu-Li’s D* and F* tests also support the (Supplementary Table 3) Tajima’s D values of hemagglutinin nucleotide sequences. The statistically significant Fu-Li’s D* and F* signify that there is positive selection affecting the changes on the molecule. This combined with the nonsynonymous TD values further show that there is positive selection on the hemagglutinin of influenza.

Furthermore, the inferred exponential growth rate values of hemagglutinin also shows that the negative Tajima’s D values are also a result of the growing population of the virus. The lower exponential growth rate, combined with more positive nonsynonymous Tajima's D values are reflected in the overall Tajima's D values. Here it is seen that the Asia, Europe, North America growth rates are many times over the smaller continents. This is also reflected in the overall Tajima's D values (Figure 3.26)

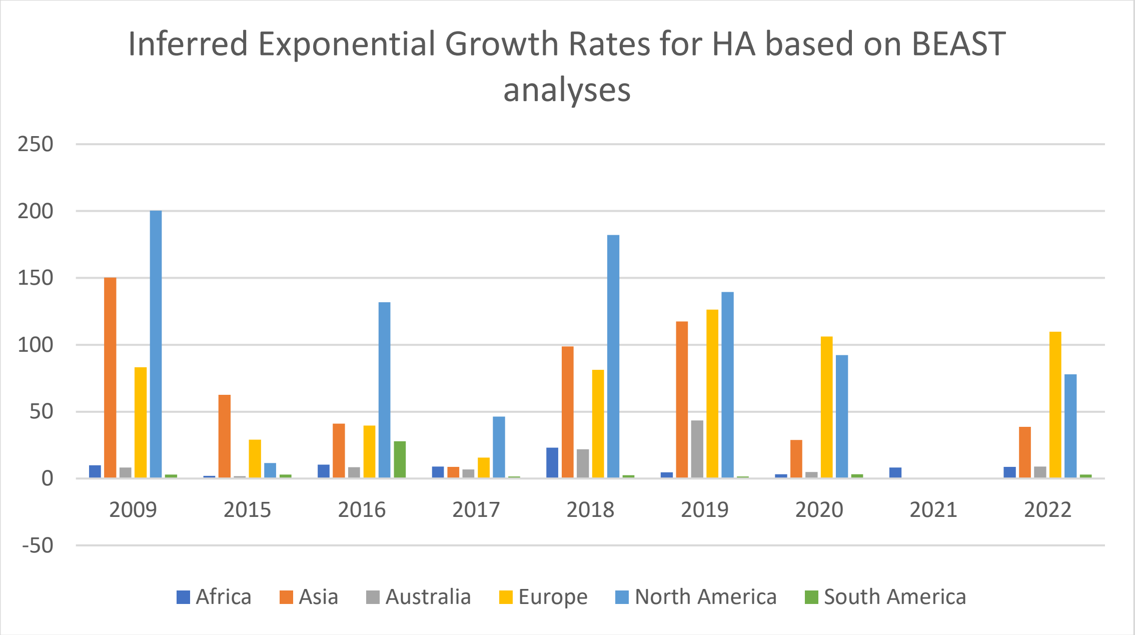


Figure 3.33 The inferred exponential growth rate of H1N1 from the BEAST analyses using hemagglutinin nucleotide sequences throughout the years.

Looking at the neuraminidase nucleotide sequences, overall Tajima's D results show that even in non-pandemic years, the positive selection acting on the nucleotide sequence is still there. Apart from 2015 and 2021, the TD values are not significantly different (Wilcoxon test P > 0.1) from the pandemic year of 2009. This signals that the positive selection on non-pandemic years apart from 2015 and 2021 may have similar positive selection and the growth rates for those years might be close to the pandemic year.

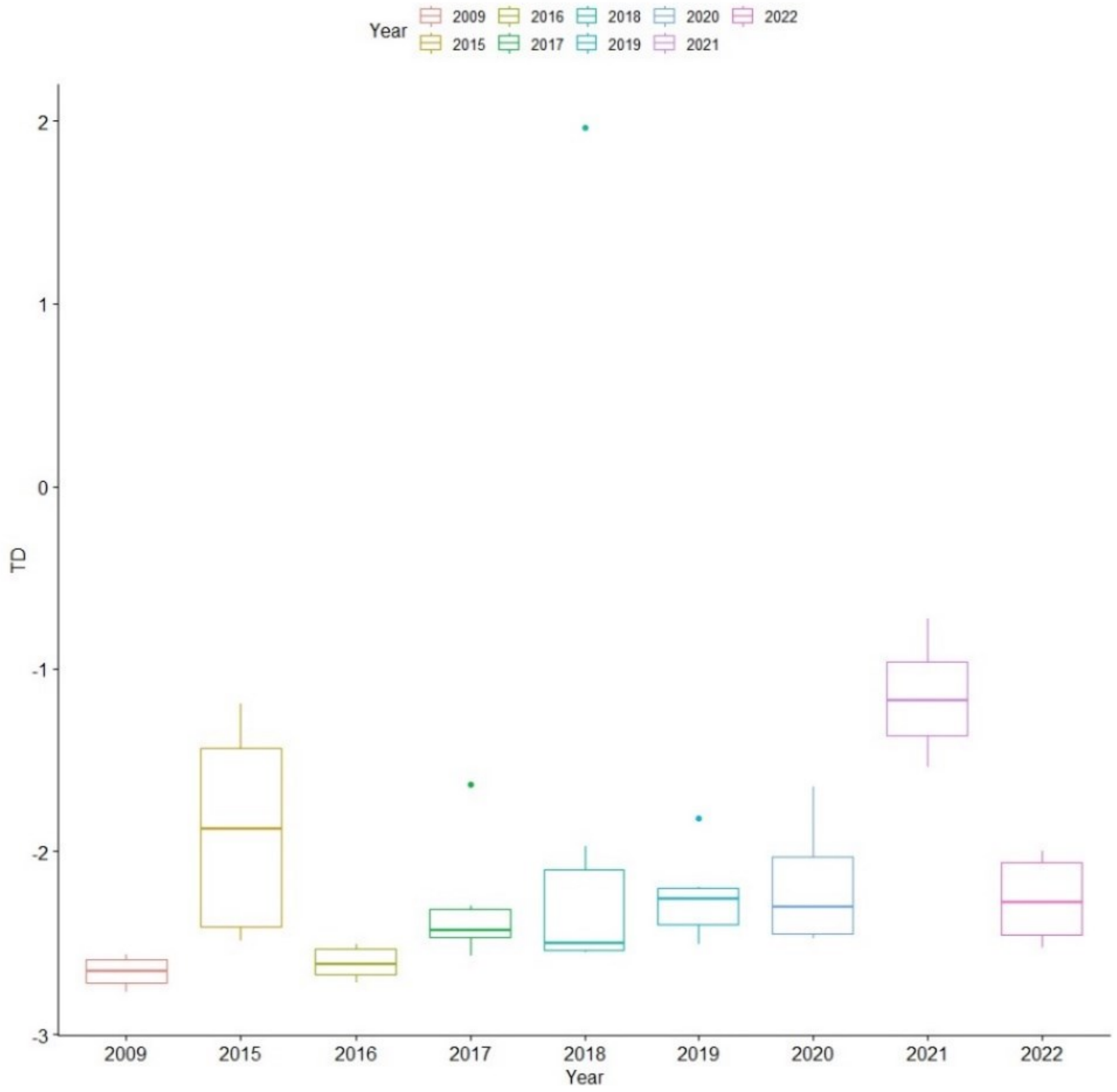


Figure 3.34 The overall Tajima’s D test results of neuraminidase nucleotide sequences separated by year from 2009 to 2022.

The location-based data shows that the bigger continents (Asia, Europe, North America) have more negative overall TD values compared to the smaller continents (Africa, South America, Australia). This is indicative of positive selection not being as strong and the population growth of these continents not being as high as the other ones. To get further information about positive selection and population growth, nonsynonymous TD values and inferred exponential growth rates are analysed.

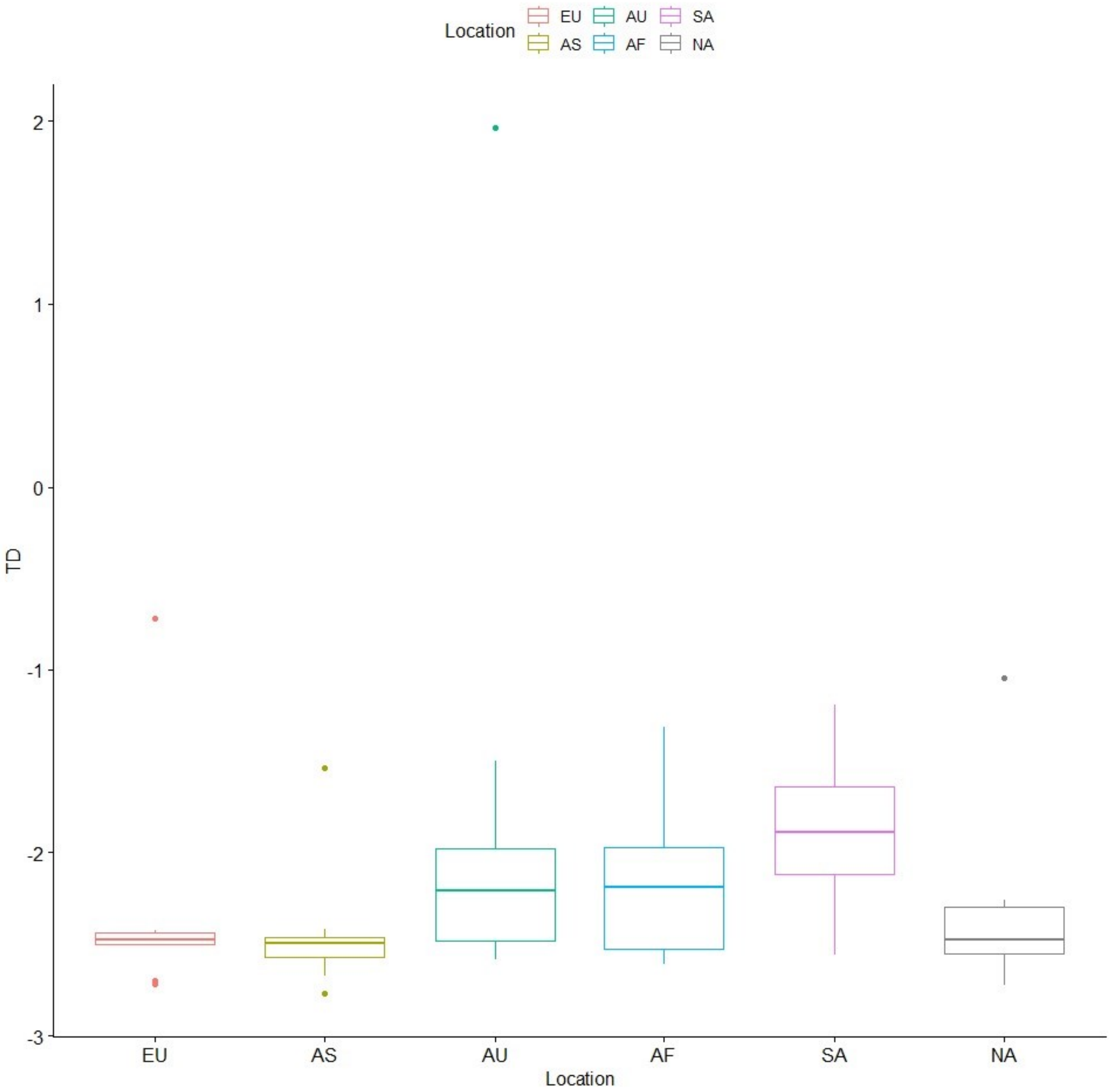


Figure 3.35 The overall Tajima’s D test results of neuraminidase nucleotide sequences separated by continents.

The nonsynonymous Tajima’s D values for neuraminidase nucleotide sequences show that pandemic year of 2009 has the most negative value, which indicates that the positive selection acting in that year is greater compared to the non-pandemic years. After 2009, the nonsynonymous TD values are more positive, which indicates the positive selection on the neuraminidase sequences are relaxed. In 2021, the nonsynonymous TD values are significantly more positive (Wilcoxon test $P < 0.05$), meaning the least amount of positive selection is acting on this year. After year 2021, the nonsynonymous TD results are more negative, indicating the positive selection acting on the neuraminidase sequences are returning.

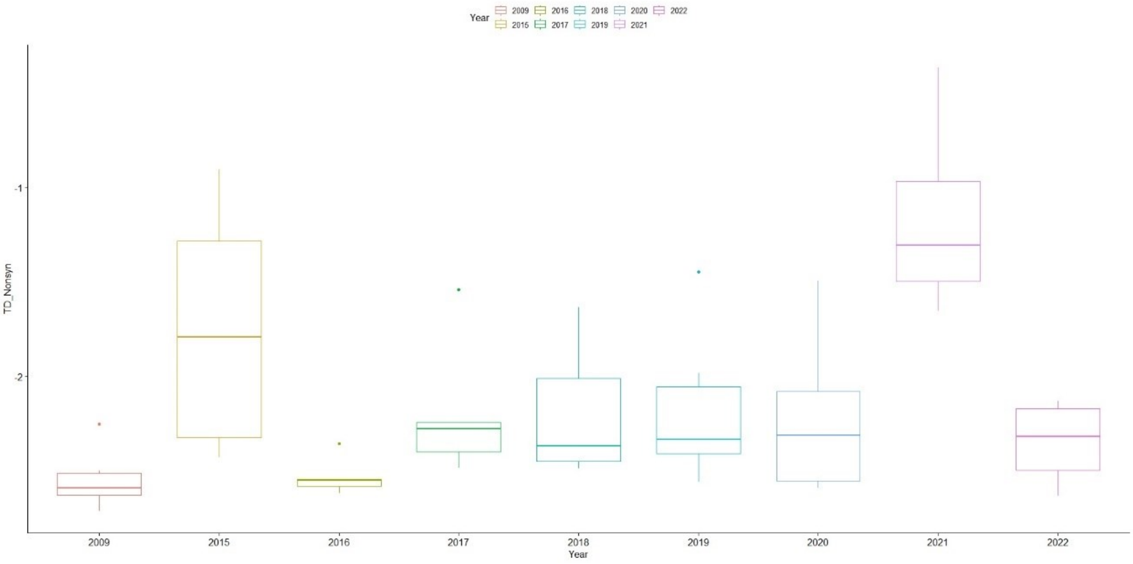


Figure 3.36 The nonsynonymous Tajima's D test results of neuraminidase nucleotide sequences separated by years from 2009 to 2022.

The nonsynonymous TD results indicate that Europe, Asia, and North America are significantly more negative (Wilcoxon test $P < 0.05$), therefore these continents are more positively selected compared to the other continents. Because the nonsynonymous TD values give primarily information about selection, synonymous TD values and inference of growth rates are important to get a clearer picture about TD values.

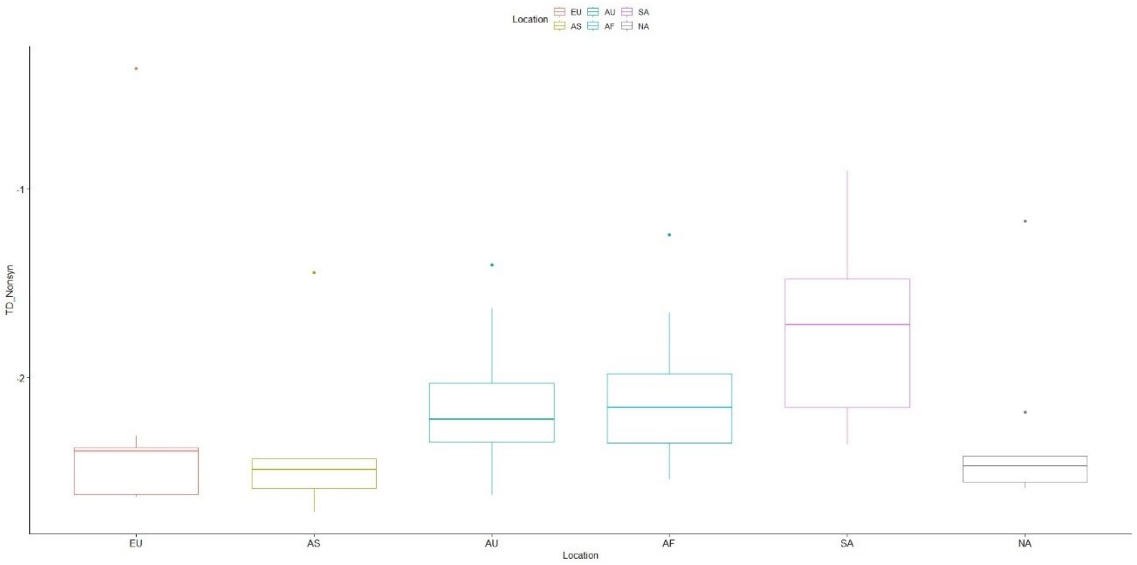


Figure 3.37 The nonsynonymous Tajima's D test results of neuraminidase nucleotide sequences separated by continents.

The synonymous Tajima's D values gives information about the population growth because synonymous changes are assumed to be silent and therefore do not affect the selection. The 2009 values are the most negative, which indicates that the growth rate in 2009 is the highest out of all years. The statistically significant difference between 2021 and all other years except 2015 (Wilcoxon test $P < 0.05$) shows that in 2021, the population growth was lesser.

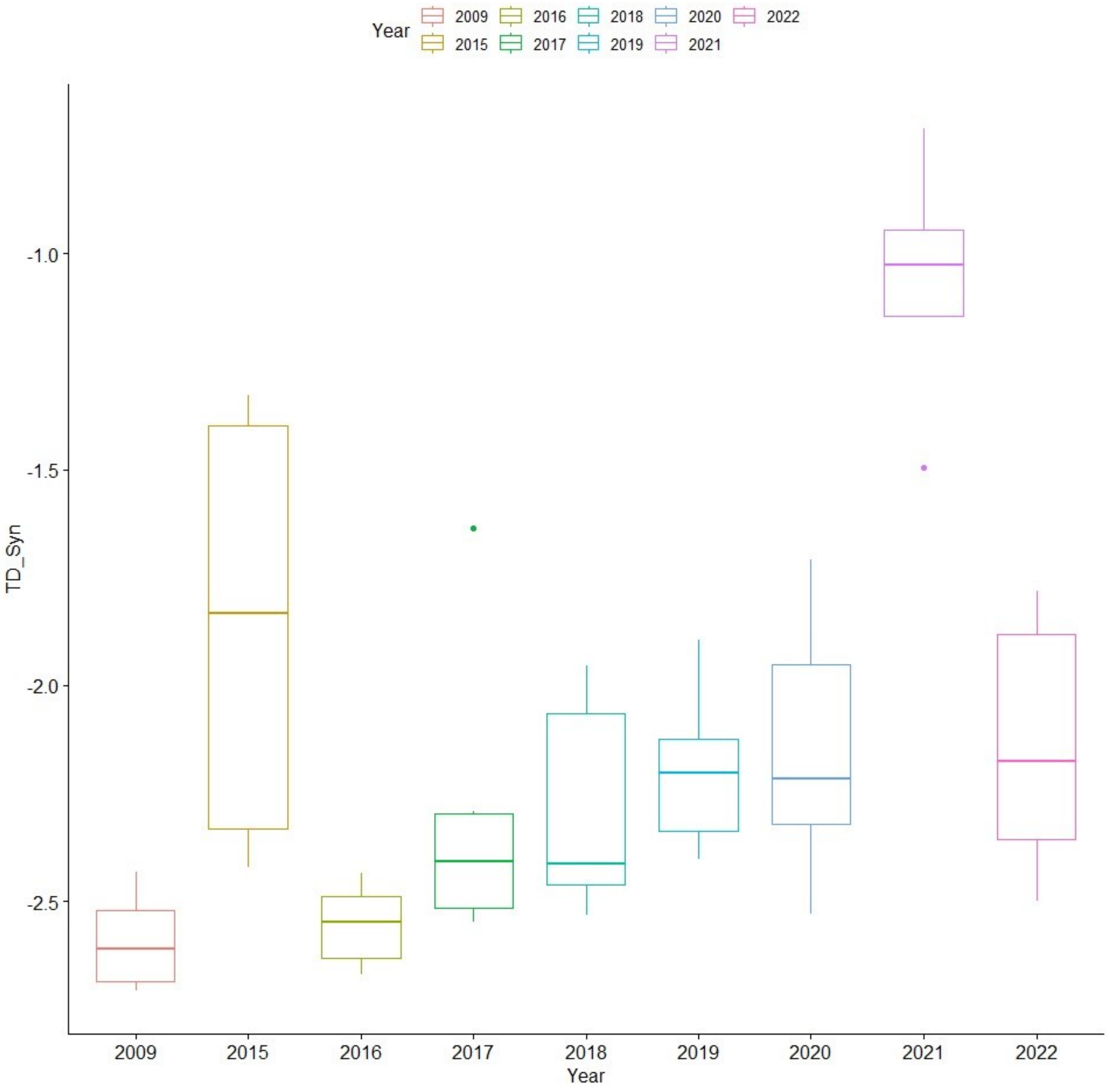


Figure 3.38 The synonymous Tajima's D test results of neuraminidase nucleotide sequences separated by years from 2009 to 2022.

There are no statistically significant differences between the big continents (Europe, Asia, North America) and the smaller continents (Africa, Australia, South America) (Wilcoxon test $P > 0.1$). Because of this, how much the population grows for each continent cannot be said.

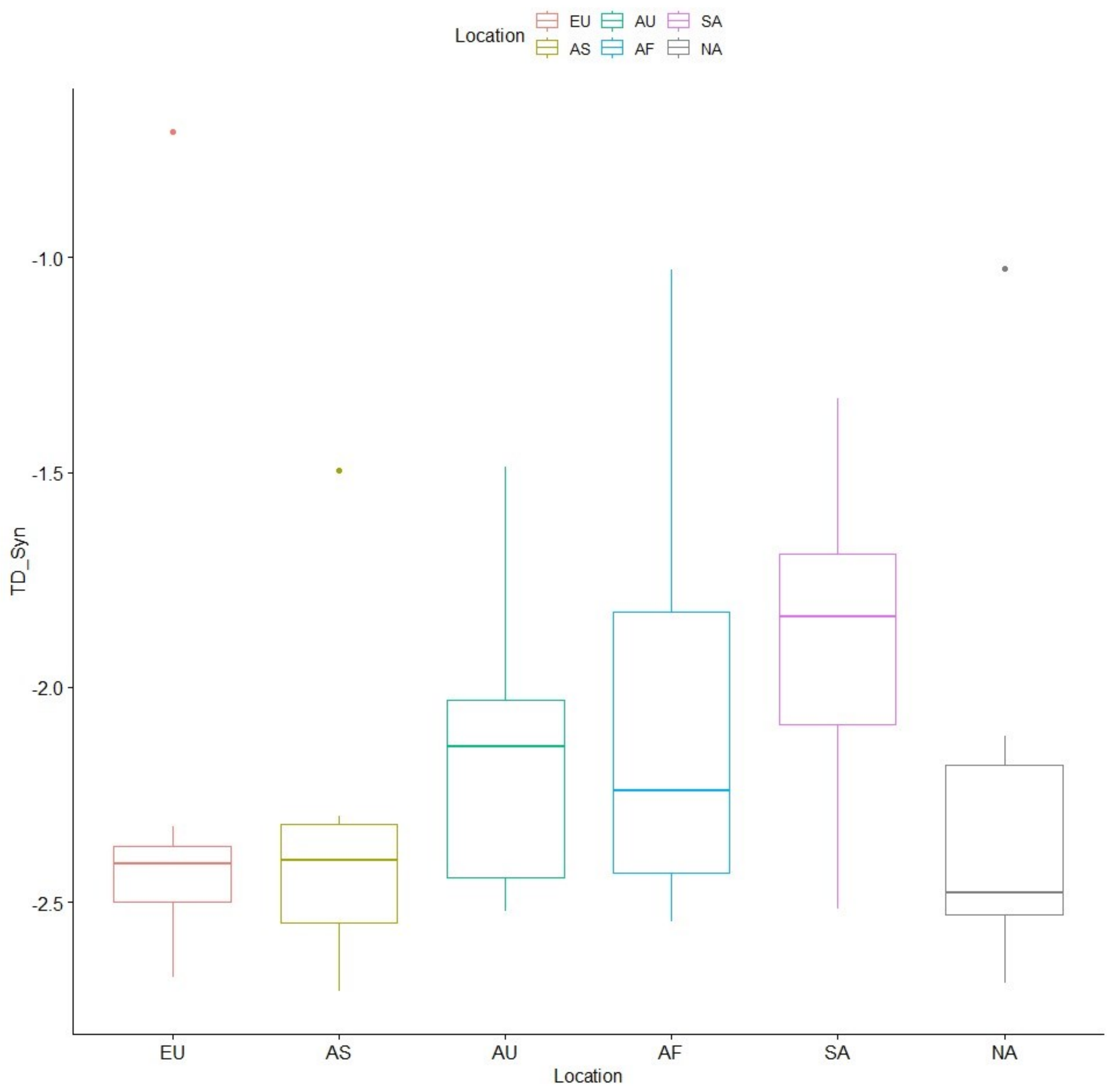


Figure 3.39 The synonymous Tajima’s D test results of neuraminidase nucleotide sequences separated by continents.

Fu-Li’s D* and F* tests give more specific information about the positive selection effects on the calculation of Tajima’s D values. For neuraminidase, the Fu-Li’s D* and F* tests support the TD values (Supplementary Table 4). This, with the findings of nonsynonymous Tajima’s D values show that there is indeed a positive selection happening on the neuraminidase of influenza.

Alongside that, the inferred exponential growth rate values of neuraminidase also shows that the negative Tajima’s D values are also a result of the growing population of the virus. The lower exponential growth rate, combined with more positive nonsynonymous Tajima's D values are reflected in the overall Tajima's D values. Here it is seen that the Asia, Europe, North America growth rates are many times over the smaller continents. This is also reflected in the overall Tajima's D values (Figure 3.32)

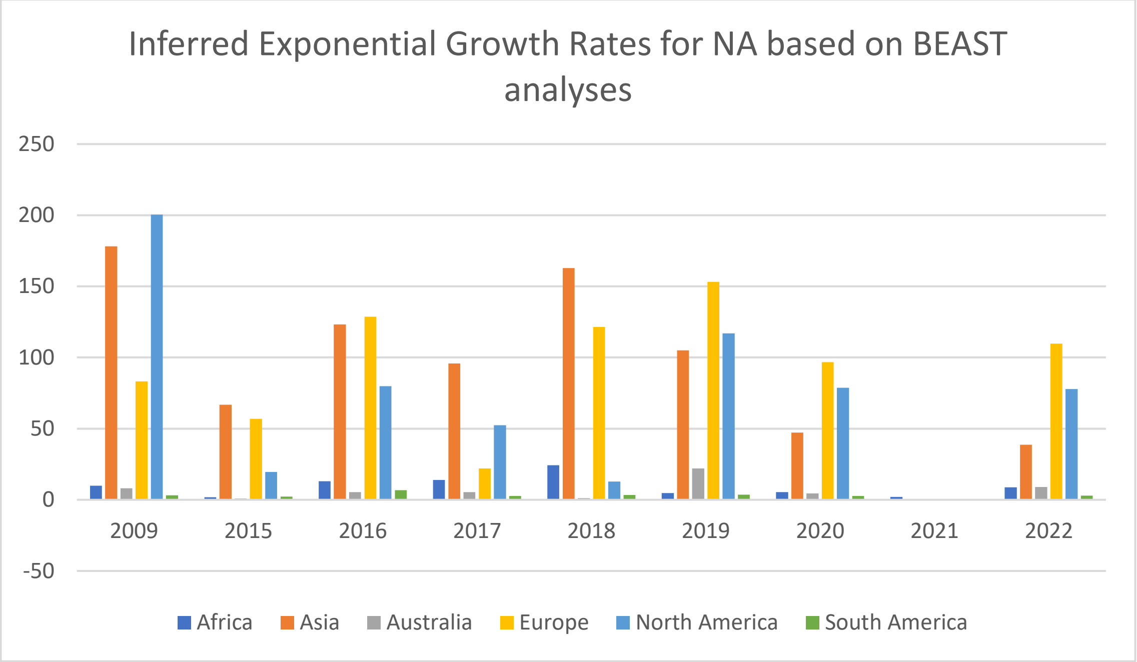


Figure 3.40 The inferred exponential growth rate of H1N1 from the BEAST analyses using neuraminidase nucleotide sequences throughout the years.

The neutrality tests for NA show a similar result to the HA of the 2009 pandemic H1N1 influenza. Almost all of the Tajima's D values are negative with statistical significance. Similar to HA, 2009 sequences scored the most negative values for Tajima's D tests. The results of Tajima's D tests and Fu-Li's D* and F* results align, with the latter having statistical significance just like Tajima's D tests. Comparing 2009 and 2022, 2009 sequences are more negative (Supplemental Table 3, 4).

Allele frequency tests using an outgroup sequence that can be aligned to the pandemic 2009 pandemic H1N1 sequences can give further information on the nature of selection on the viral sequences because these tests use the outgroup viral sequence to polarize observed mutations as ancestral or derived.

For the hemagglutinin outgroup selection, human H2N5 HA, human H5N1 HA, swine H1N1 HA, human influenza B HA were tried. Other than the H5N1 HA sequences, none of the HA sequences could be aligned properly with the human H1N1 sequences used in this study and therefore H5N1 HA was selected. For neuraminidase, human H3N2, human influenza B NA, swine H1N1 NA sequences were tried. Swine H1N1 NA nucleotide sequences were the best aligned with the human H1N1 NA sequences and therefore swine H1N1 NA was selected as the outgroup for human H1N1 NA analyses.

For HA, human H5N1 HA nucleotide sequence was used as an outgroup for all of the years and continents. Fu-Li's D and F tests show that there is still a large amount of rare variants among the 2009 pandemic H1N1 with tests having statistical significance (Supplementary Table 3). Fay and Wu's H and normalized F tests were also negative, indicating an excess of derived rare variants. When compared to 2022, the 2009 Fu and Li's D and F tests show a more negative result, in line with the Fu-Li's D* and F* tests that are performed without an outgroup (Supplementary Table 3, 5).

The NA neutrality tests with an outgroup were done using NA nucleotide sequence of swine H1N1 as the outgroup. The Fu-Li's D and F tests for almost all of the samples are negative and have significance (Supplementary Table 6). This signals that there is an excess amount of rare variants in the NA sequences.

The McDonald-Kreitman test gives information about the number of fixed and polymorphic synonymous and nonsynonymous changes on the sequences, as well as the neutrality index about the sequences.

The McDonald-Kreitman test also gives information about the neutrality index (NI) and the alpha value. The neutrality index gives information about how much the sequences derive from being neutrally selected. D_s = Synonymous fixed substitutions, D_n = nonsynonymous fixed substitutions, P_n = polymorphic synonymous substitutions, P_n = nonsynonymous polymorphic substitutions. The neutrality index

$$NI = (P_n/P_s / D_n/D_s)$$

A neutrality index greater than 1 (NI > 1) indicates a negative selection on the population which manifests itself as an excessive amount of polymorphism. A neutrality index smaller than 1 (NI < 1) is indicative of positive selection because of a greater amount of nonsilent divergence in the population.

Alpha value, which is formulized by;

$$\alpha = 1 – (D_sP_n/D_nP_s)$$

gives information about the proportion of substitutions driven by positive selection. The alpha value can range between - ∞ and 1. The closer alpha value is to 1, greater the proportion of substitution driven by positive selection in the population.

Direction of selection (DoS) is a derivative of McDonald-Kreitman test, its formula being:

$$DoS = D_n/(D_n + D_s) – P_n(P_n + P_s)$$

the direction of selection test gives information about the selection on the population even with scarce amount of data. 0 means a neutral direction, negative values indicate a negative selection and positive values indicate a positive selection [42].

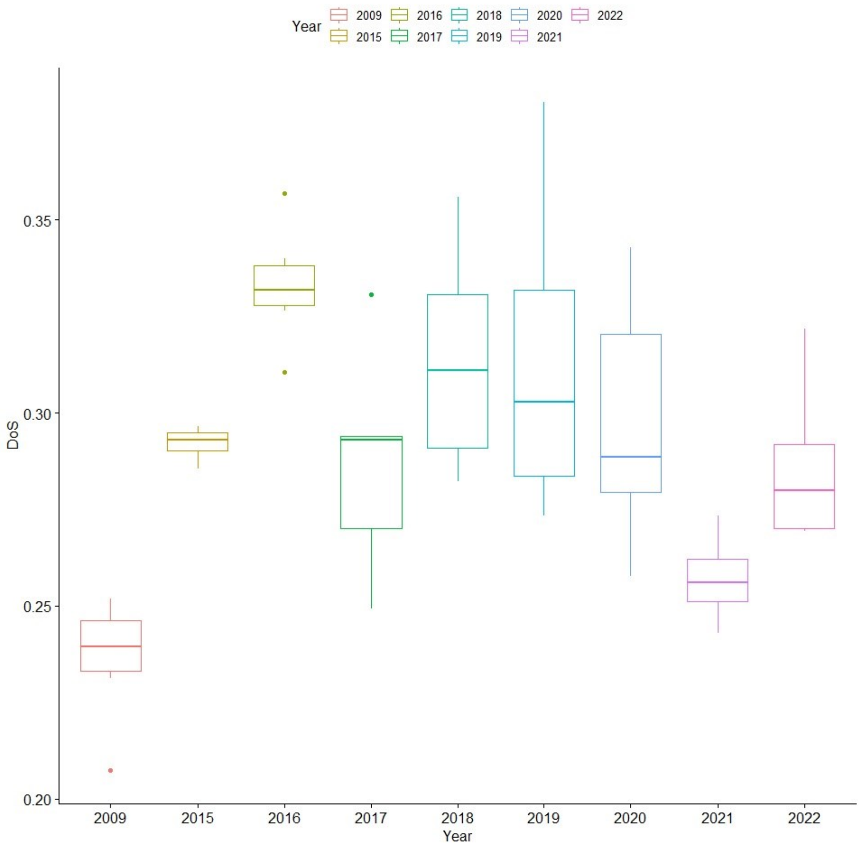


Figure 3.41 Direction of selection (DoS) test results of hemagglutinin nucleotide sequences by using the outgroup human H5N1 HA nucleotide sequences.

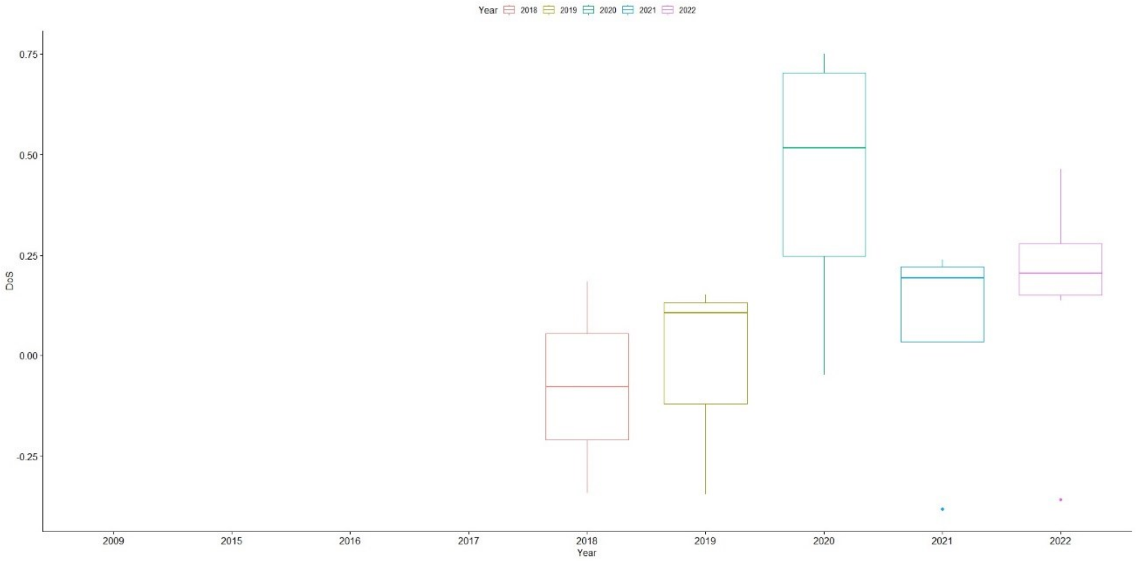


Figure 3.42 Direction of selection (DoS) test results of neuraminidase nucleotide sequences by using the outgroup swine H1N1 NA nucleotide sequences.

The McDonald-Kreitman tests for HA show that there is a slightly positive selection on the populations. The neutrality index of 2009 Europe samples are the highest in the dataset, in line with other tests about the selection being positive. The lowest neutrality index value belongs to the 2019 North America samples. The neutrality index gets lower as time approaches to 2021, however after 2021 there is an increase to the neutrality index and nonsynonymous changes (Supplementary Table 5).

Comparing the 2009 and 2022 values, it is seen that the neutrality index is higher in 2009 values, which indicates that there was a lesser positive selection on pandemic sequences in the pandemic year compared to the post SARS-CoV-2 pandemic year of 2022.

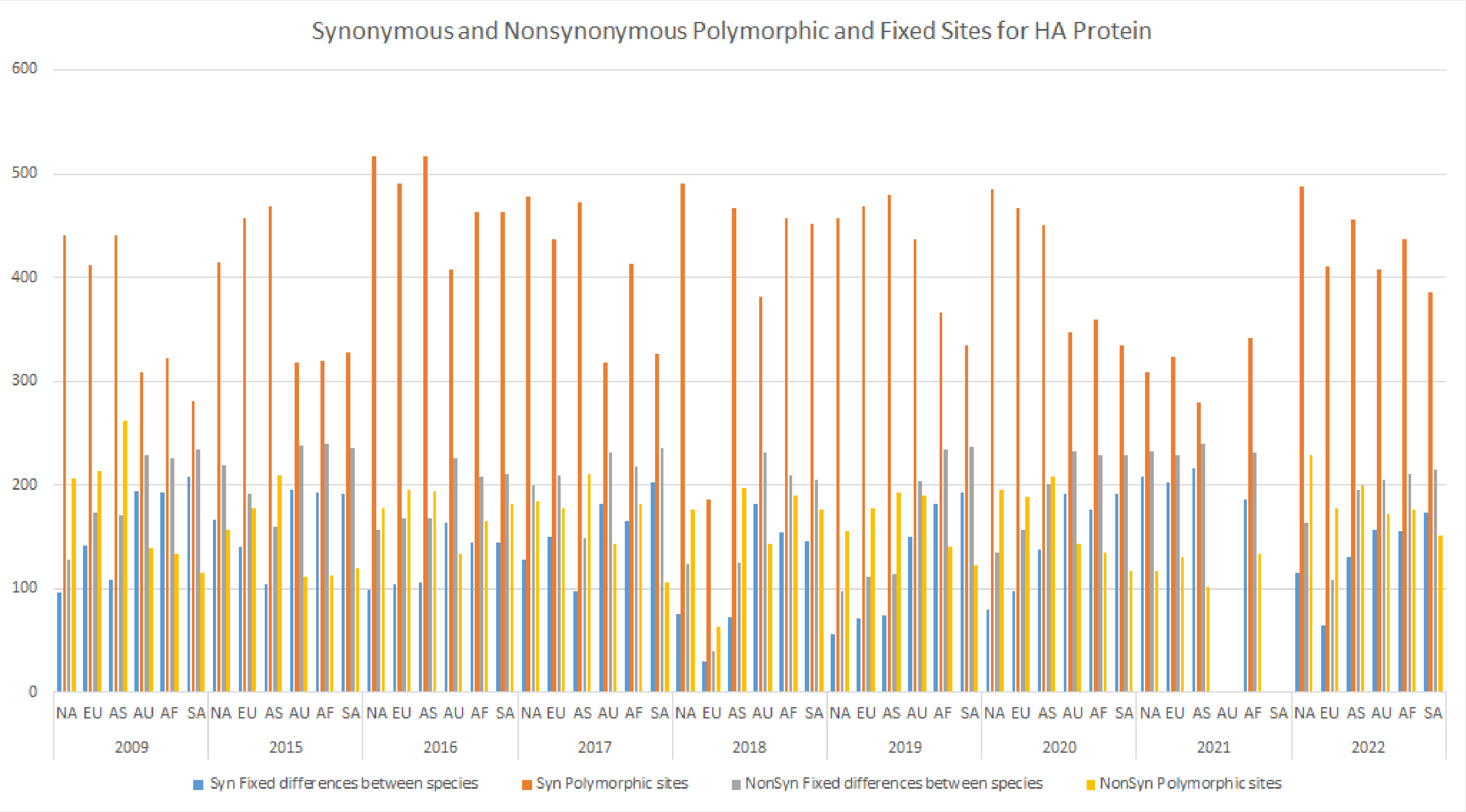


Figure 3.43 McDonald-Kreitman test results of hemagglutinin nucleotide sequences by using the outgroup human H5N1 HA nucleotide sequences.

The NA samples have very different results compared to the HA results. The McDonald-Kreitman tests show no fixed differences between swine N1 and human N1 until year 2018 (Supplementary Table 6). Because there were no fixed differences between swine H1 and human H1 subtypes, the neutrality index and degree of selection cannot be calculated.

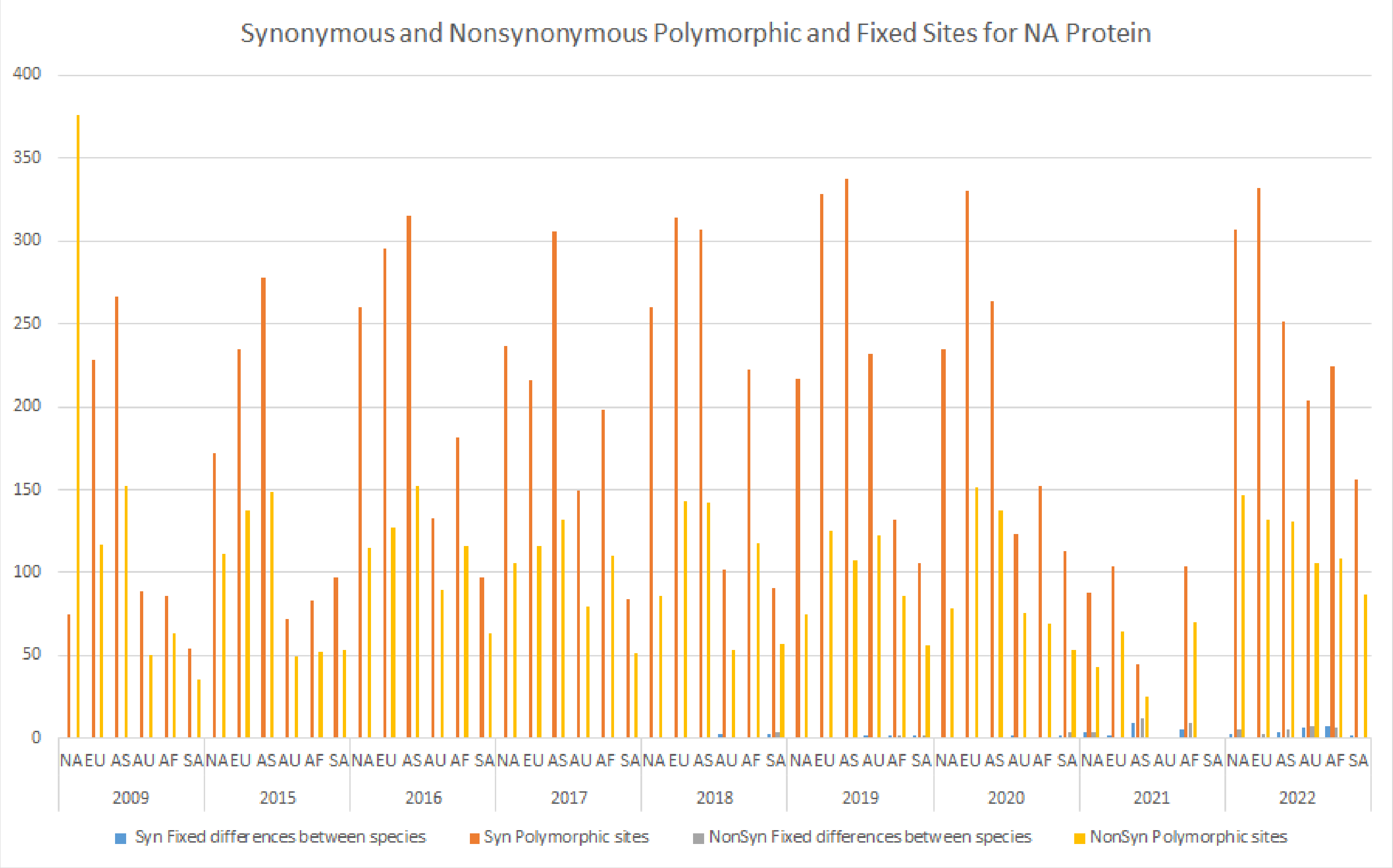


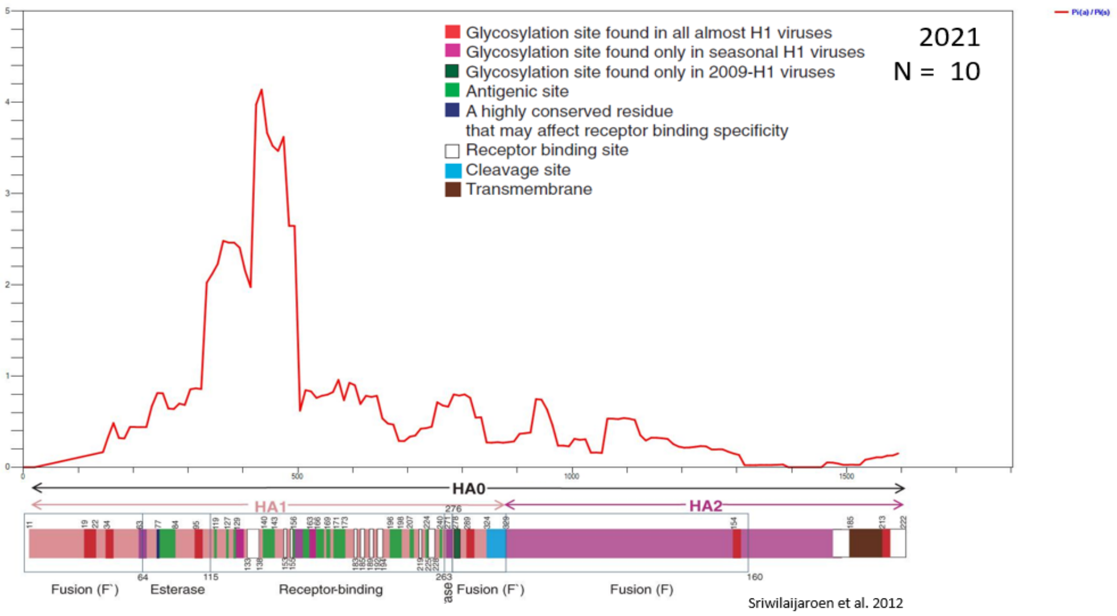
Figure 3.44 McDonald-Kreitman test results of neuraminidase nucleotide sequences by using the outgroup swine H1N1 NA nucleotide sequences.

3.3 Sliding Window Analyses of Amino Acid Changing (Nonsynonymous) Nucleotide Diversity

The sliding window analyses show information about the nonsynonymous (Pi(a)/Pi(s)) changes happening on the amino acid sequences of the hemagglutinin and neuraminidase. The information here gives more specific meaning to the nonsynonymous changes happening on the molecules. Knowing which part of the amino acid sequence part corresponds to which domain and function, information can be inferred as the most rapidly changing and conserved parts of the molecules.

3.3.1 Sliding Window Analyses of Amino Acid Changing (Nonsynonymous) Nucleotide Diversity of 2009 pandemic influenza H1N1 HA Nucleotide Sequences

The 2021 Asia HA sliding window analysis show high peaks on the receptor binding site and antigenic site of the receptor binding site in HA1. HA2 part shows much less change. The changes on the receptor binding site may be helping the HA bind itself to sialic acid in a more efficient way. The changes in antigenic site help may be helping the influenza virus evade the immune system with novel and unrecognizable antigenic patterns. However, the positive selection for year 2021 is not strong based on the nucleotide diversity and Tajima's D values therefore the nature of these changes cannot be inferred based on the sliding window analysis.



Sriwilaijaroen et al. 2012

Figure 3.45 Sliding window analysis of hemagglutinin nucleotide sequences for Asia continent for year 2021.

The sliding window analysis for 2022 Asia continent show a stark difference compared to the 2021 Asia sliding window analysis. While there are many peaks, the rate of changes is very little compared to the 2021 Asia analysis. There are many peaks, and even in the more conserved part of the HA, HA2, there are multiple peaks. The biggest peaks in HA1 correspond to the antigenic sites on the receptor binding domain of hemagglutinin. These changes may be helping with the immune escape. The changes on the cleavage site of the fusion (F’) part of the HA1 may help with the conformational change of HA after it is bound to sialic acid.

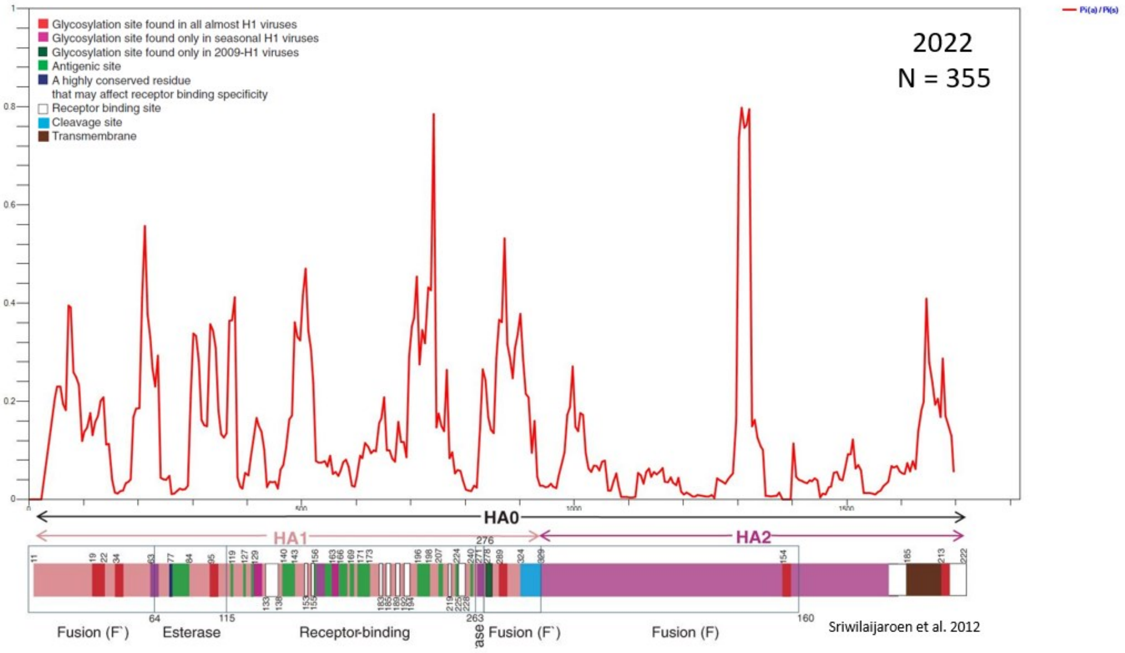


Figure 3.46 Sliding window analysis of hemagglutinin nucleotide sequences for Asia continent for year 2022.

The sliding window analysis of HA for North America 2021 show lower amount of nonsynonymous changes compared to the sliding window analysis of HA for Asia 2021. The highest peak of Pi(a)/Pi(s) for North America 2021 is 0.8 meanwhile, for Asia 2021 it is more than 4.0. This indicates while there are changes happening on the HA for North America for year 2021, the changes are fewer compared to Asia. The biggest peaks on the HA molecule for North America for 2021 are on the receptor binding domain of HA1. The antigenic sites and receptor binding sites on the receptor binding domain show the most changes. These changes may be helping with immune evasion and more optimal binding to the receptor (sialic acid) respectively.

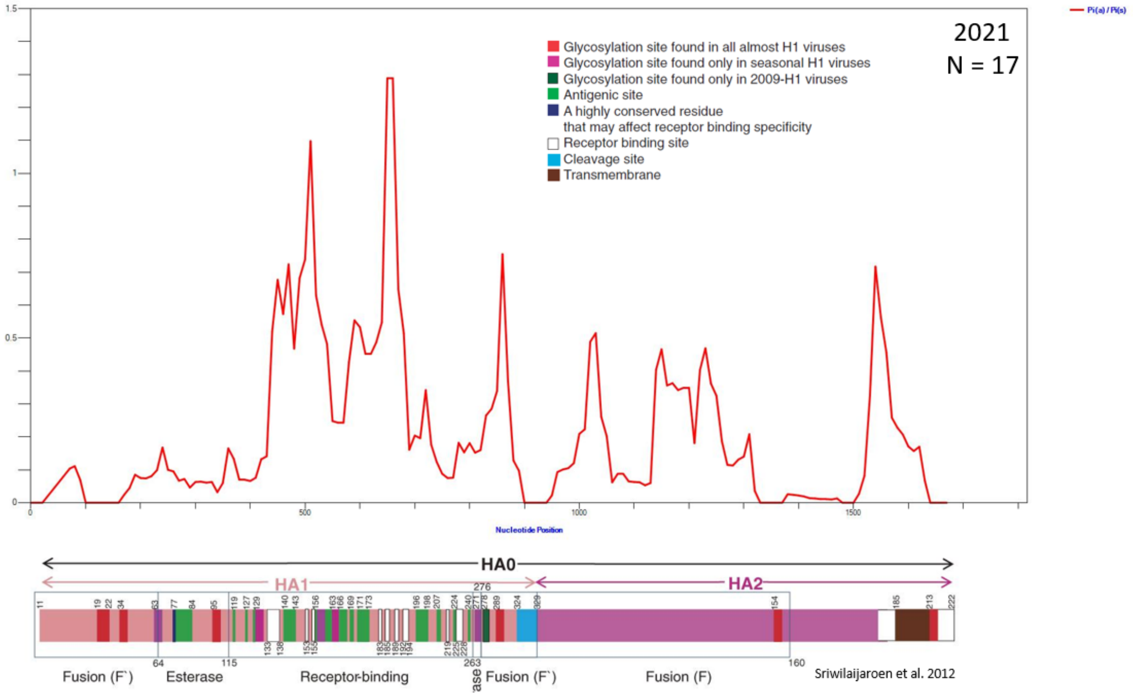


Figure 3.47 Sliding window analysis of hemagglutinin nucleotide sequences for North America continent for year 2021.

The sliding window analysis for HA for North America 2022 shows that the biggest peak is in the fusion (F’) domain of the HA1. The cleavage site is important for the conformational change of hemagglutinin after its binding with the sialic acid, which helps with the fusing the host and viral cell membranes. A secondary big peak is in the antigenic sites of the receptor binding domain. The changes on the antigenic sites may be helping the hemagglutinin evade the host immune systems. The peak on the receptor binding sites on the receptor binding domain indicates that the nonsynonymous changes may be helping with more optimal binding to hemagglutinin’s receptor. These changes are presumed to be positive because of the nucleotide diversity, Tajima’s D, BEAST results all indicating that there are positive selections on the sequences.

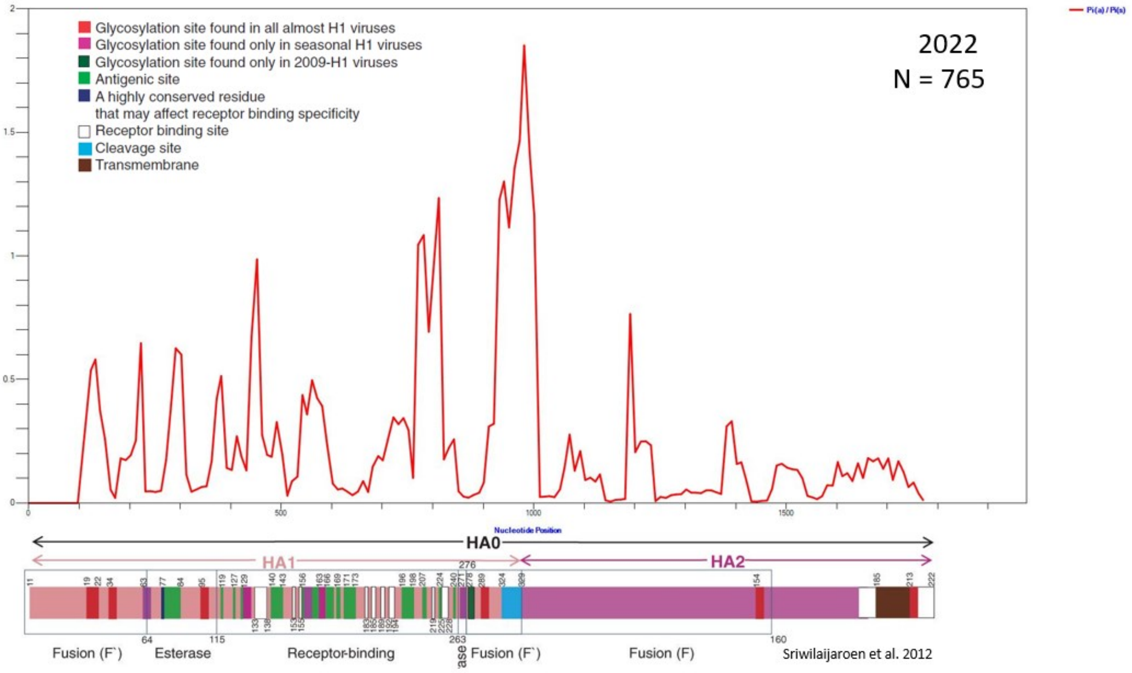
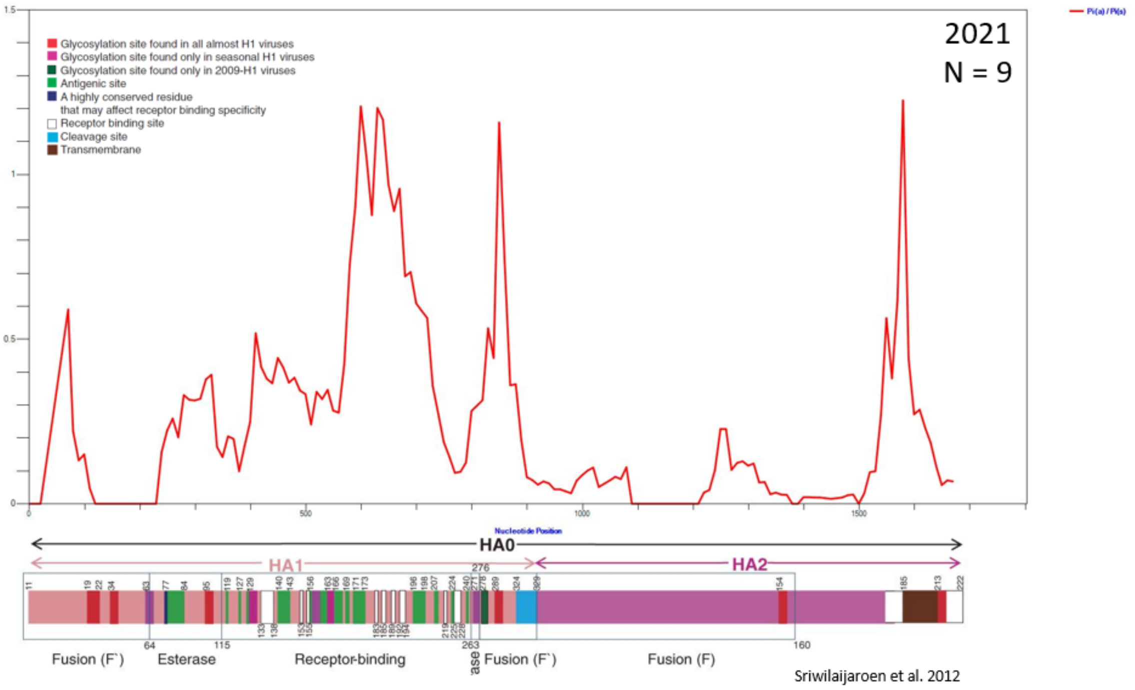


Figure 3.48 Sliding window analysis of hemagglutinin nucleotide sequences for North America continent for year 2022.

The rate of nonsynonymous changes on the hemagglutinin sequences of 2021 European continent is similar to the 2021 North American rates, however still lower than the 2021 Asia rates. There are four big peaks on the HA sequences, and they are on the antigenic sites and receptor binding sites of the receptor binding domain of HA1, the glycosylation site on the fusion (F’) domain of HA1, and the transmembrane domain of the fusion domain (F) of HA2.

The nonsynonymous changes occurring on the antigenic sites on the receptor binding domain may be helping with the immune evasion of influenza. Because the antigenic sites are recognized by the host immune system, changes occurring in these sites may make these sites unrecognizable for the host immune system, allowing the influenza virus to go undetected. The changes on the receptor binding sites of the receptor binding domain may be helping with optimizing hemagglutinin’s binding to the sialic acid receptor.

The changes on the glycosylation site of fusion domain (F’) of HA1 may be affecting how hemagglutinin changes conformation after its initial binding with sialic acid. The glycosylation plays a role in the modification of the protein, such as folding correctly, stabilizing the protein. The changes on the transmembrane region of the fusion domain (F) in HA2 may be helping with the structural integrity and the stability of the hemagglutinin by docking the protein more safely and stabile into the viral membrane of influenza.



Sriwilaijaroen et al. 2012

Figure 3.49 Sliding window analysis of hemagglutinin nucleotide sequences for European continent for year 2021.

The rate of the nonsynonymous changes on the hemagglutinin molecule has almost doubled compared to the 2021 European sequences. The biggest peak is seen on the glycosylation site on the fusion (F') domain, with a smaller peak seen on the cleavage site on the fusion (F') domain on the HA1. Another big peak is observed on the antigenic sites on the receptor binding domain in HA1. Some smaller peaks are observed on the antigenic sites and the receptor binding sites on the upstream of the big peak on the receptor binding domain.

The changes on the glycosylation site found on the fusion (F') domain on the HA1 may be helping with more efficient conformational change of the hemagglutinin after its binding with sialic acid. The folding and structure of the fusion domain may also be affected positively with these changes. The changes on the cleave site could be also helping with more optimal change of confirmation of HA.

The changes on the antigenic sites of the receptor binding domain may be helpful with better immune evasion of influenza from the host immune system. The changes on the receptor binding site on the receptor binding domain may be helping with optimizing the hemagglutinin-sialic acid binding.

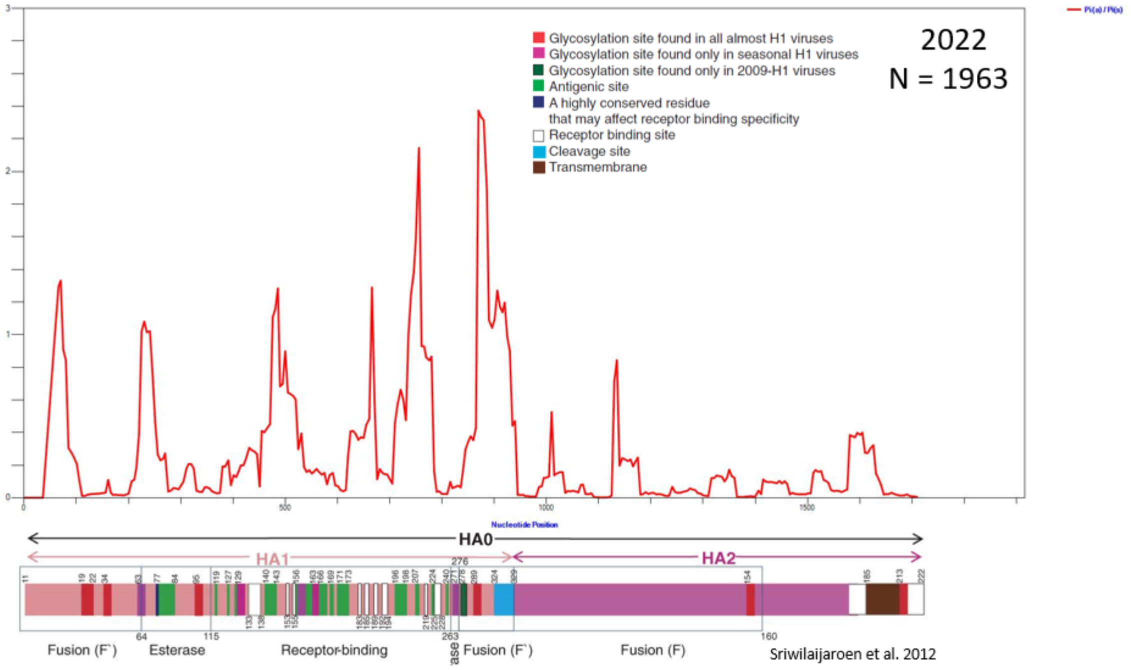


Figure 3.50 Sliding window analysis of hemagglutinin nucleotide sequences for Europe continent for year 2022.

There is only one big peak on the hemagglutinin sliding window analysis of the 2022 Australia nucleotide sequences which is on the antigenic site of the receptor binding domain of HA1. While there is only peak on the sliding window analysis, the Pi(a)/Pi(s) rate is 5 to 10 times that of Asia, North America, Europe. This suggests that this spot on the Australia 2022 sequences is a hotspot for nonsynonymous changes. The spot is on the antigenic site, which suggests that the changes occurring here may be helpful with influenza's immune evasion of human immune system.

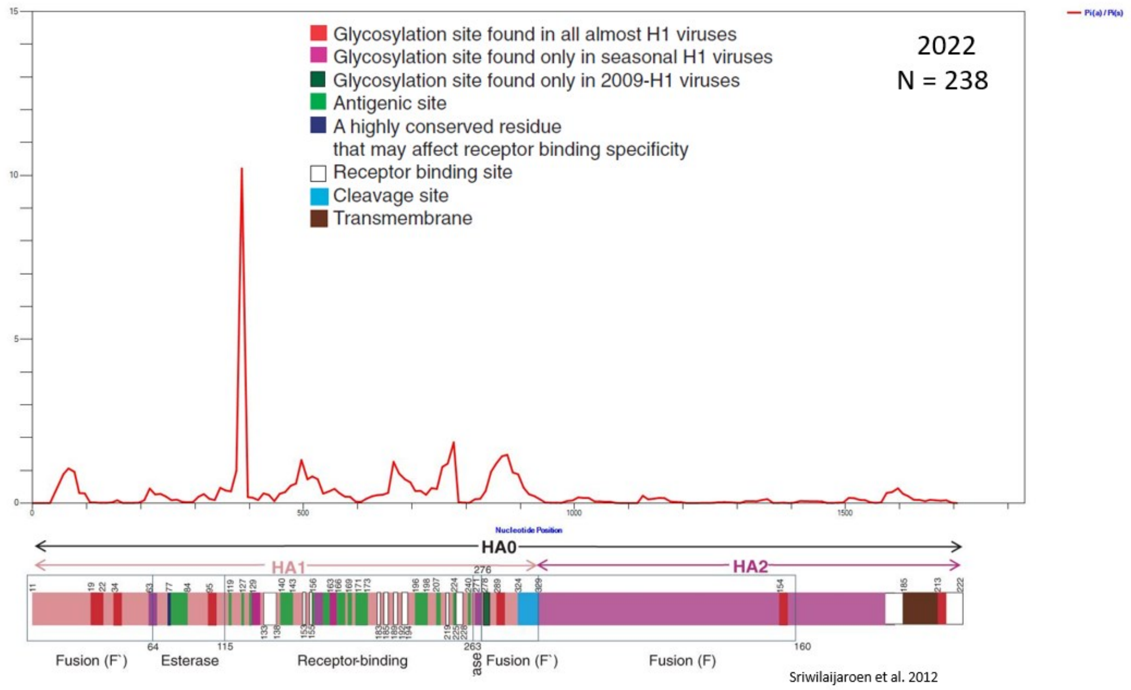
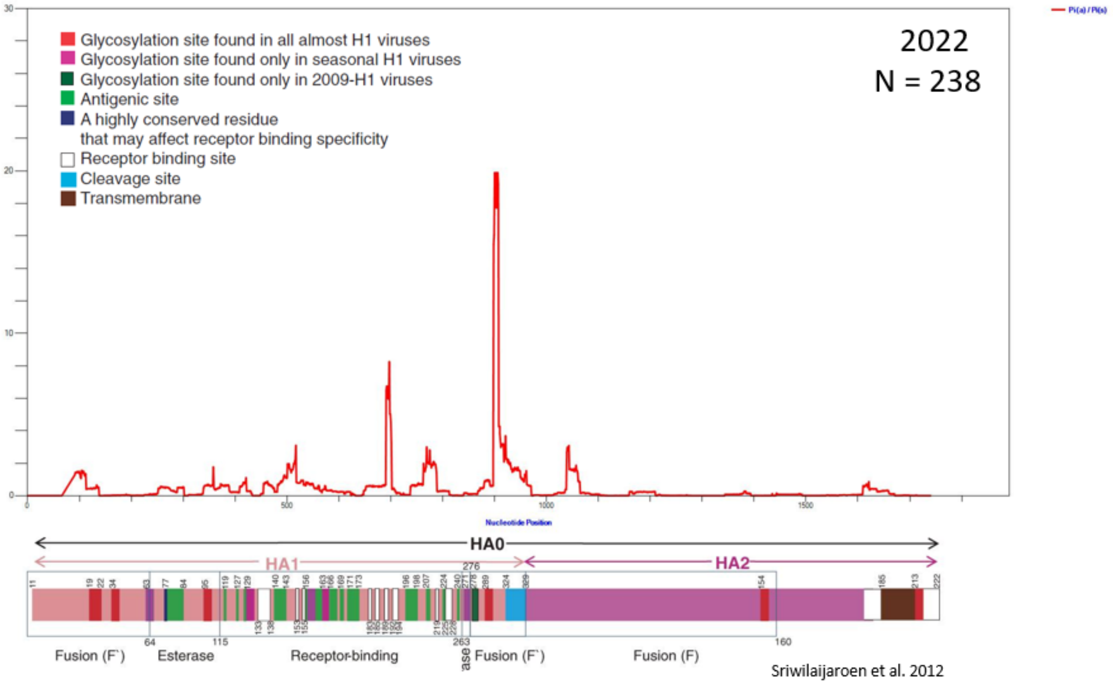


Figure 3.51 Sliding window analysis of hemagglutinin nucleotide sequences for Australia continent for year 2022.

The sliding window analysis of hemagglutinin nucleotide sequences for South America for year 2022 show a big peak on the glycosylation site on the fusion (F') domain of HA1, with a smaller peak on the receptor binding site on the receptor binding domain of HA1. The Pi(a)/Pi(s) rate of the biggest peak is 20, which is almost two times more than the 2022 Australia's biggest peak, and 10 to 20 times more than the Asia, North America, Europe peaks. This suggest that the nonsynonymous changes happening on these peaks are even greater than the other continents listed.

The peak on the glycosylation site on the fusion (F') domain of HA1 may be optimizing the folding, conformational change, stability of the fusion domain as the glycosylation helps with correct confirmation of proteins and the function of the protein. The secondary smaller peak on the receptor binding site of the receptor binding domain may have positive effects on the binding of hemagglutinin to the sialic acid, which is its receptor.



Sriwilaijaroen et al. 2012

Figure 3.52 Sliding window analysis of hemagglutinin nucleotide sequences for South America continent for year 2022.

The sliding window analysis result for the hemagglutinin nucleotide sequences for Africa continent for year 2022 show the greatest Pi(a)/Pi(s) rates out of all the hemagglutinin sliding window analyses. The biggest peak is on the fusion(F) domain of the HA2. There are smaller clusters of peaks on the glycosylation site and cleavage site of the fusion (F') domain in HA1, and on the antigenic site of the receptor binding domain in HA1.

The nonsynonymous changes on the fusion (F) domain may be affecting the stability of the fusion (F) domain of the hemagglutinin, as well as the efficiency of the conformational change of the hemagglutinin after its initial binding with sialic acid receptor. The nonsynonymous changes on the fusion (F') domain of HA1 may be of similar nature. The changes on the glycosylation sites could be helping with correct folding, function, stability of the hemagglutinin while the changes on the cleavage site may be helping with more optimal conformational change of the hemagglutinin once it binds to the sialic acid receptors.

The changes on the antigenic sites may help with the immune evasion of influenza virus.

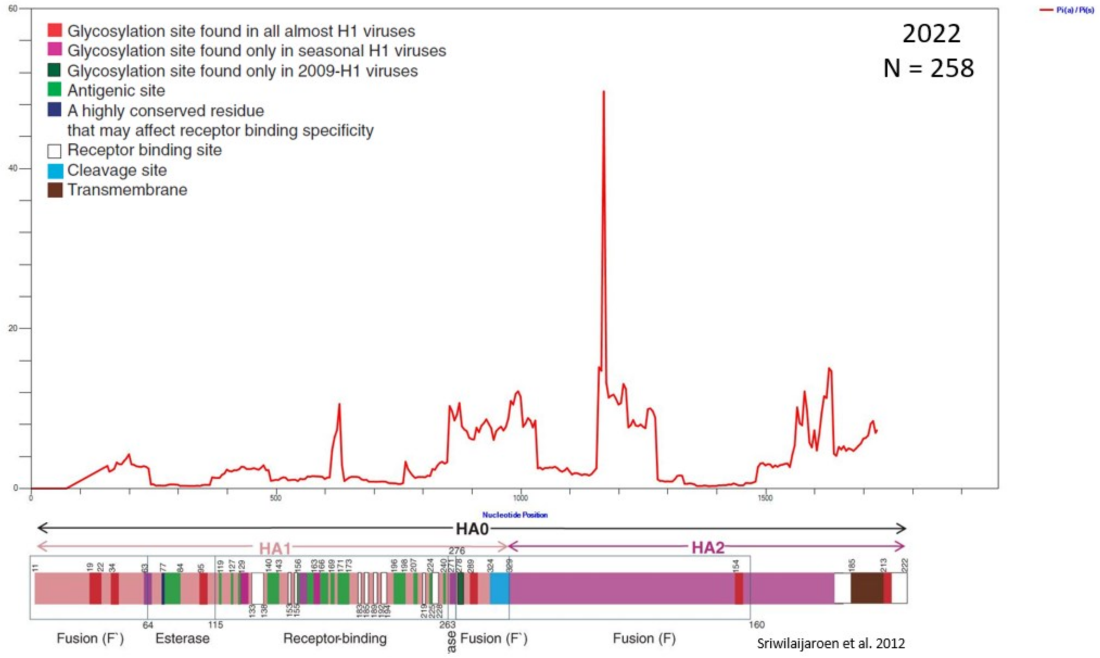


Figure 3.53 Sliding window analysis of hemagglutinin nucleotide sequences for Africa continent for year 2021.

Most of the nonsynonymous changes seen on the hemagglutinin nucleotide sequences reside within the HA1 section of the molecule which houses the fusion, esterase, and receptor-binding domains. The peaks are where the most nonsynonymous changes happen and most of these peaks are seen on the antigenic sites on various domains and the receptor binding sites on the receptor binding domain.

The changes on the antigenic sites may help with survival of the influenza by making the antigenic sites unrecognizable by the host immune system. The immune evasion can help positively by making the virus easier to survive and spread due to being undetectable by the host immunity.

Changes on the receptor binding sites may help with better binding of hemagglutinin to the sialic acid receptor. The binding of hemagglutinin to sialic acid receptor is the first step of entrance to the host cell, therefore the changes happening on these sites may help with more efficient and secure binding, increasing the virulence of influenza.

The cleavage site on the fusion (F’) domain is another site with many peaks, and the fusion domain of hemagglutinin helps with the conformational change of hemagglutinin after binding to sialic acid receptor and the following fusion of viral and host membranes. The changes happening on these sites may help with more efficient conformational changes.

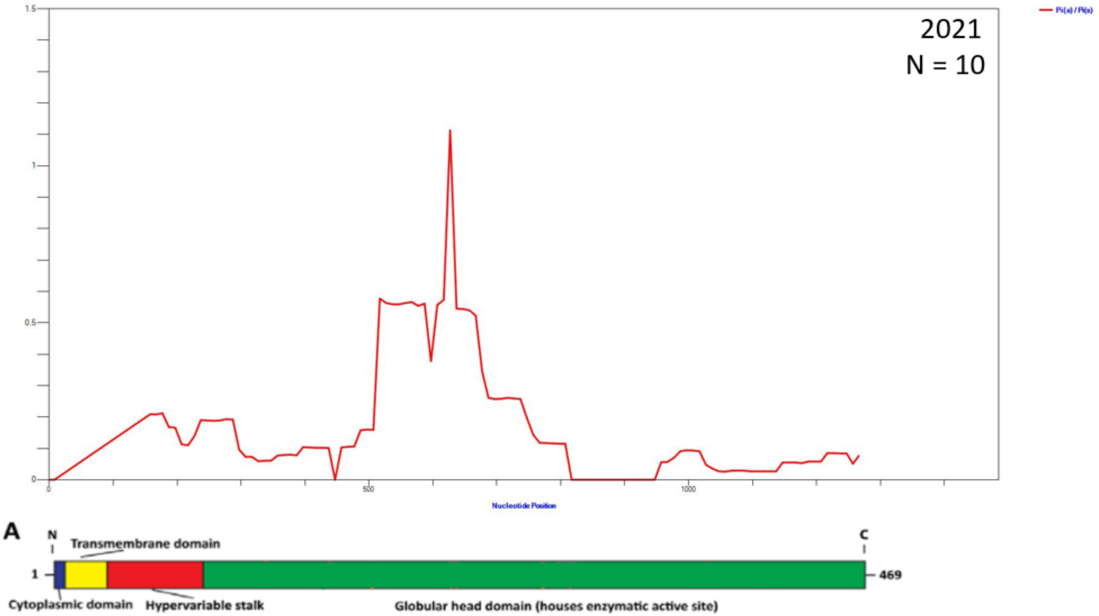
HA2, which houses the fusion (F) site of hemagglutinin is much more conserved compared to HA1, and there were very few changes in comparison to number of changes in HA1. By looking at the sliding window analysis results, it can be said that the HA2 of hemagglutinin is very much conserved in an evolutionary sense.

In the bigger continents (Asia, North America, Europe) the sliding window analysis for 2022 sequences all showed many different peaks happening over the sequences, however when looked at the smaller continents (Australia, South America, Africa) the 2022 sequences only have one big peak. This means that the hemagglutinin in bigger continents has more than one hot spot that has many nonsynonymous changes while the hemagglutinin in smaller continents has one spot that has many nonsynonymous changes. One reason why this may have happened is because the exponential growth rates are higher in the bigger continents, which means the influenza has spread to a larger amount of hosts, which means that the hemagglutinin had more chances of replications and therefore mutations and nonsynonymous changes.

3.3.2 Sliding Window Analyses of Amino Acid Changing (Nonsynonymous) Nucleotide Diversity of 2009 pandemic influenza H1N1 NA Nucleotide Sequences

The sliding window analysis of neuraminidase nucleotide sequences of Asia for year 2021 show a big peak in the globular head domain of the neuraminidase, as well as some smaller peaks on the transmembrane domain and hypervariable stalk domain of the molecule. The changes on the globular head domain, which contains the enzymatic active site, may positively affect the neuraminidase activity of cleaving the sialic acid-hemagglutinin link. While not as big, the peaks on the hypervariable stalk domain of neuraminidase signal nonsynonymous changes also happening there. The changes on the hypervariable stalk are very important to the functionality of the neuraminidase because the optimal stalk height is paramount to the survival of the virus. Longer than normal stalk will interfere with the entrance of the virus into the host cell by preventing the initial hemagglutinin-sialic acid binding, which is required to facilitate viral cycle. Shorter than normal stalk will be unable to cleave the bonds between hemagglutinin and sialic acid once a new viral particle is formed. This failure to cleave the bond will prevent the newly formed viral particle from leaving the host cell, therefore preventing the virus from spreading after the initial infection.

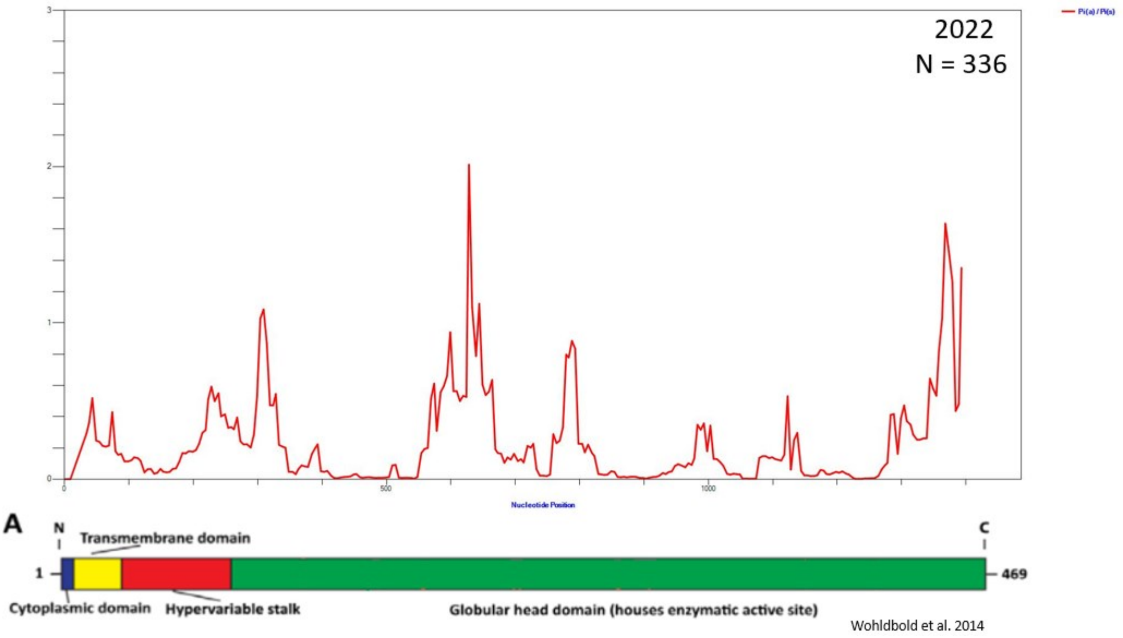
Because the changes on the stalk are very important to the survival of the virus, the nonsynonymous changes that are seen on the stalk domain may be optimizing the stalk height for optimal neuraminidase enzymatic activity. The stability of neuraminidase may be also affected by the nonsynonymous changes on both the transmembrane domain as well as the hypervariable stalk domain.



Wohldbold et al. 2014

Figure 3.54 Sliding window analysis of neuraminidase nucleotide sequences for Asia continent for year 2021.

Compared to the nonsynonymous changes in 2021 Asian neuraminidase nucleotide sequences, 2022 Asian neuraminidase nucleotide sequences show a greater amount of nonsynonymous change with many more peaks. Most of the peaks are in the globular head domain of neuraminidase which houses the enzymatic active site. The changes on the globular head domain may be helping with optimization of the enzymatic activity of neuraminidase. The stability of the head domain may also be affected by these nonsynonymous changes. The nonsynonymous changes on the transmembrane domain of neuraminidase may be positively affecting the stability of neuraminidase.



Wohldbold et al. 2014

Figure 3.55 Sliding window analysis of neuraminidase nucleotide sequences for Asia continent for year 2022.

The sliding window analysis of neuraminidase nucleotide sequences for North America in year 2021 show a big peak around the 700th nucleotide, and smaller peaks on the hypervariable stalk domain. The changes on the globular head domain, which contains the enzymatic active site within may be positively changing the efficiency of neuraminidase enzymatic activity of cleaving the bonds between hemagglutinin and sialic acid.

The length of the stalk of neuraminidase plays a crucial role in the survival of the virus by facilitating enzymatic activity at correct rate. Neuraminidase stalk being too long will interrupt the hemagglutinin-sialic acid binding required for the entrance to the host cell whereas the neuraminidase stalk being too short will prevent newly formed viral particles from exiting the host cell because neuraminidase will not be able to cleave the bond between the newly formed hemagglutinin and host cell's sialic acid. The changes on the stalk may be optimizing the enzymatic activity of neuraminidase.

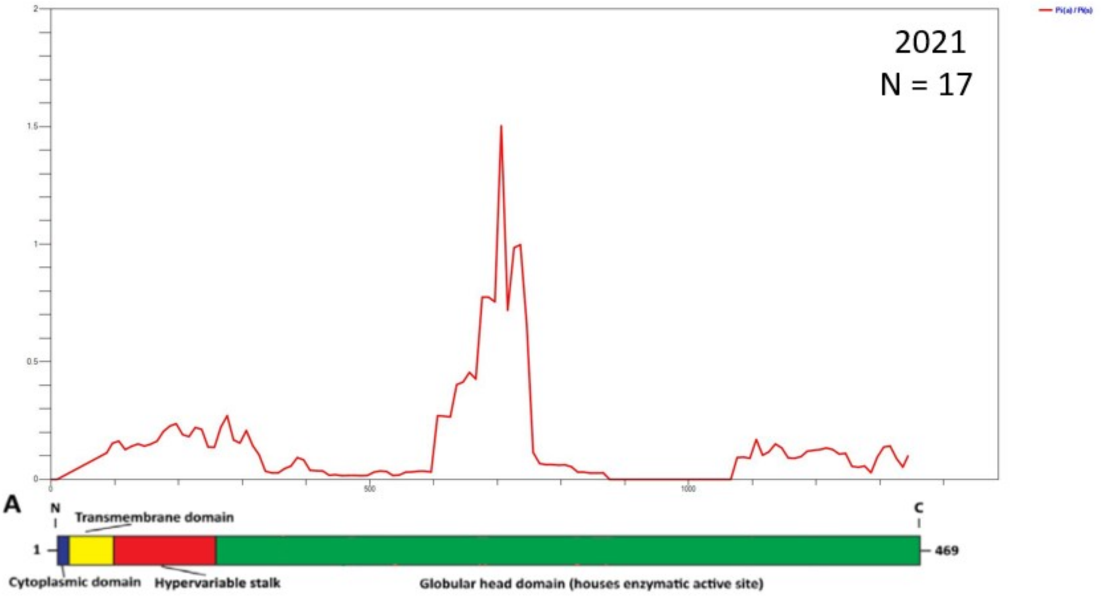
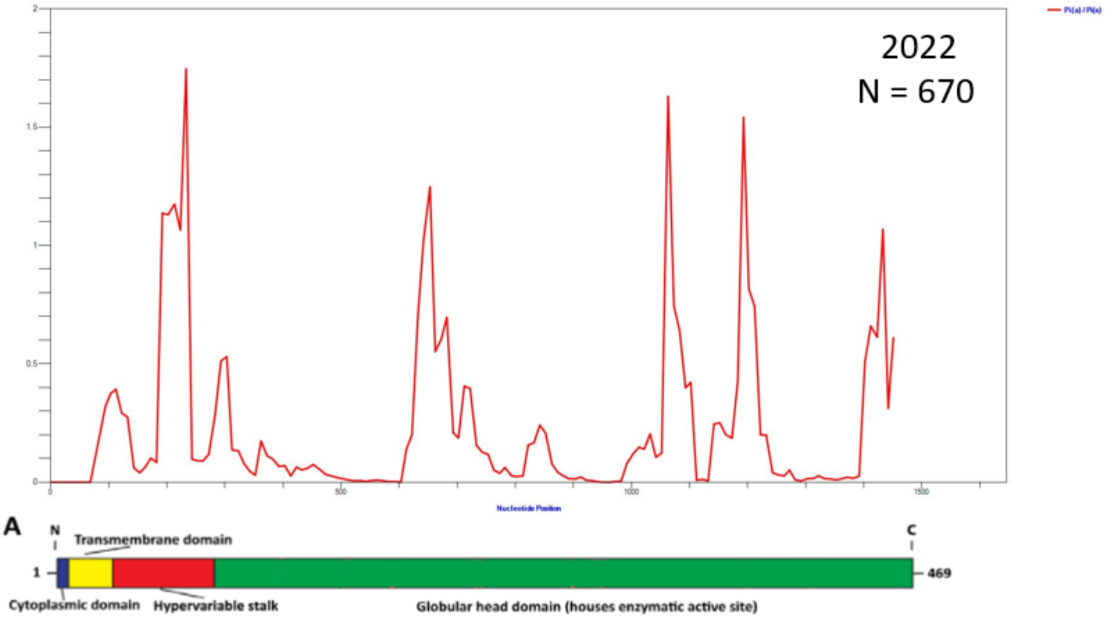


Figure 3.56 Sliding window analysis of neuraminidase nucleotide sequences for North America continent for year 2021.

Compared to the sliding window analysis of neuraminidase nucleotide sequences for North America in year 2021, the sliding window analysis showed more peaks, therefore more nonsynonymous changes in 2022. The North American neuraminidase samples in 2022 show the biggest peak on the hypervariable stalk domain among all the 2022 samples. There are four other major peaks on the globular head domain of neuraminidase which houses the enzymatic active site.

The changes on the hypervariable stalk domain may be optimizing the length and the stability of the stalk of neuraminidase. The length plays a role in optimum enzymatic activity as too long or too short stalk lengths will be detrimental to the survival of the virus.

Nonsynonymous changes on the globular head domain may be related to the enzymatic activity of neuraminidase and the stability of the head domain. Because the positive selection, these chances may be helping with optimizing the enzymatic activity of neuraminidase, improving the stability of the head domain.

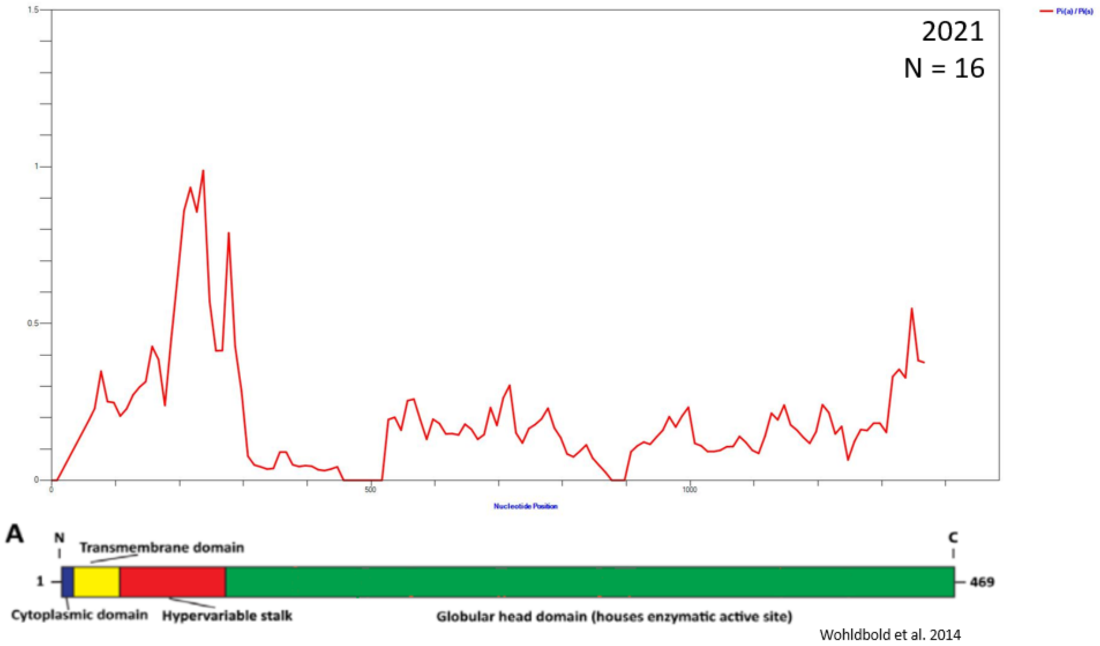


Wohldbold et al. 2014

Figure 3.57 Sliding window analysis of neuraminidase nucleotide sequences for North America continent for year 2022.

The sliding window analysis of neuraminidase nucleotide sequences for Europe continent for year 2021 show that the biggest changes are happening on the hypervariable stalk of neuraminidase, and smaller changes on the globular head domain which contains the enzymatic active site.

The changes on the hypervariable stalk may be related to the stability of the stalk part of neuraminidase as well as the optimal length of the stalk. The stability of the stalk is important for neuraminidase to have structural integrity and support for the neuraminidase. The length of the stalk of neuraminidase is really important for the proper functioning of the neuraminidase. The incorrect length of the stalk will either prevent the release of newly formed viral particles if it is too short, or it will prevent the initial entrance to the host cell if it is too long.

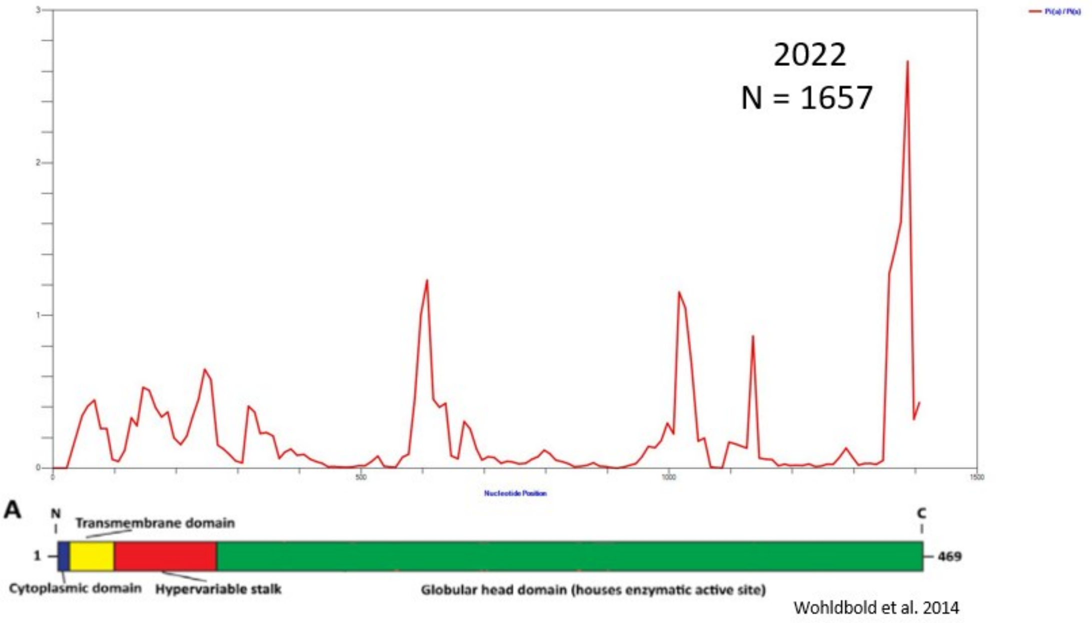


Wohldbold et al. 2014

Figure 3.58 Sliding window analysis of neuraminidase nucleotide sequences for Europe continent for year 2021.

The sliding window analysis of neuraminidase nucleotide sequences for Europe in 2022 show a big peak towards the end of the globular head domain, also throughout the globular head domain. There are smaller peaks on the transmembrane domain, and the hypervariable stalk.

The nonsynonymous changes on the globular head domain may affect the enzymatic activity of neuraminidase. The changes on enzymatic activity may help with optimizing the efficiency of neuraminidase enzymatic activity. Changes on the hypervariable stalk domain are important because the height of the neuraminidase stalk is crucial for viral function and survival. The correct height will ensure neuraminidase activity is not too much that it prevents hemagglutinin binding to sialic acid for viral entrance to host and also ensure that neuraminidase activity will not fail to cleave the hemagglutinin-sialic acid bonds once the newly viral particles are ready to leave the host cell.

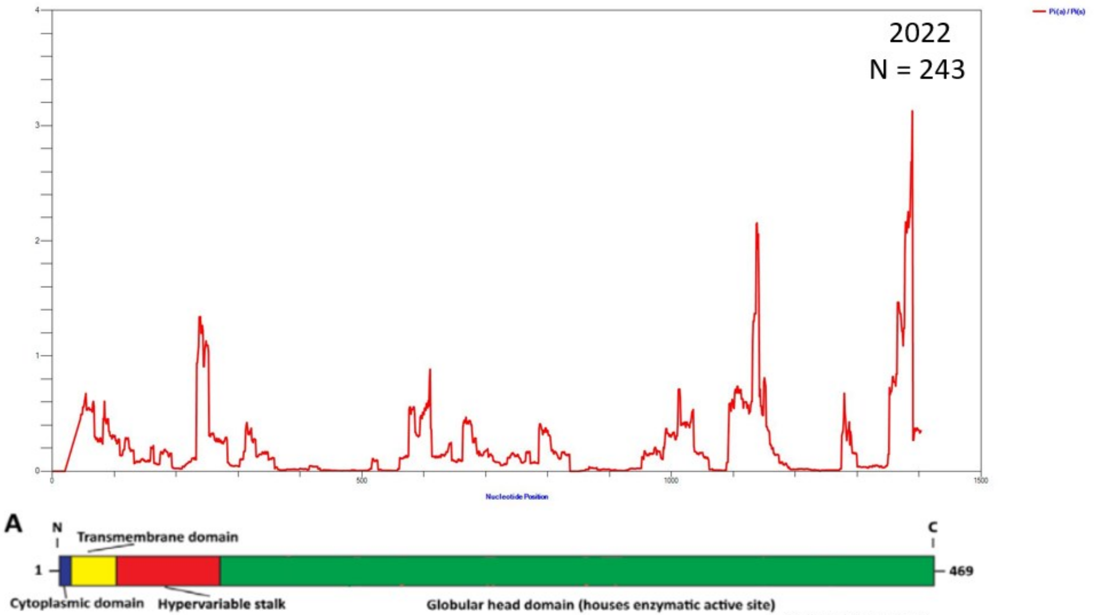


Wohldbold et al. 2014

Figure 3.59 Sliding window analysis of neuraminidase nucleotide sequences for Europe continent for year 2022.

The nonsynonymous changes on the neuraminidase nucleotide sequences for Australia for 2022 happen mostly on the globular head domain, with smaller amounts of changes happening on the transmembrane domain and the hypervariable stalk of neuraminidase.

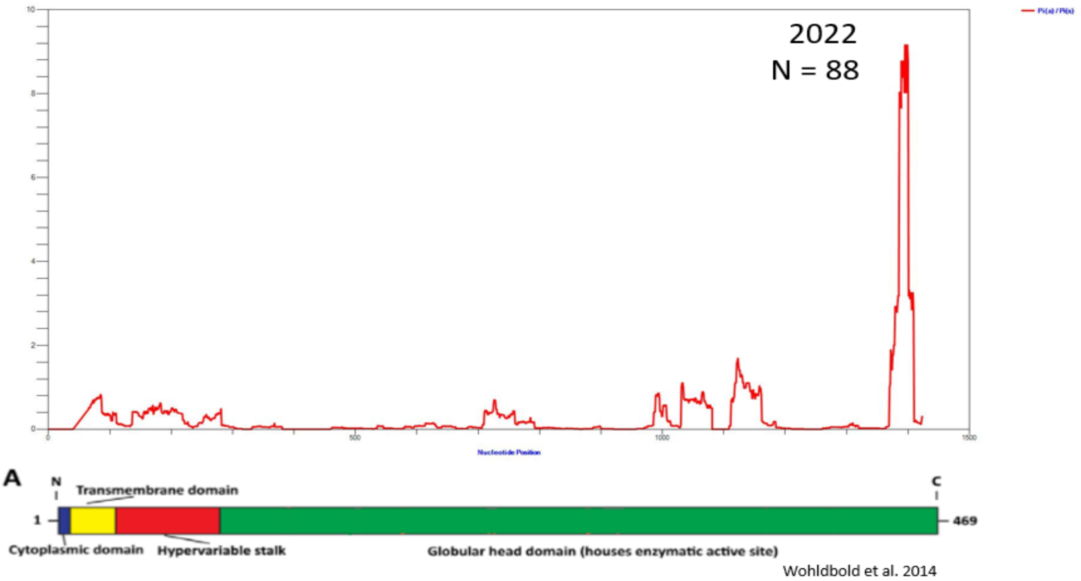
Changes on the globular head domain may affect the activity of the enzymatic site of neuraminidase, which facilitates proper release of newly formed viral particles from the host cell. These changes may stabilize the structure of the head domain, which is also important in enzymatic activity of neuraminidase. The nonsynonymous changes on the transmembrane domain may help with stability of neuraminidase stalk. The changes on the hypervariable stalk domain may also affect the stability of the stalk as well as the height of the stalk.



Wohldbold et al. 2014

Figure 3.60 Sliding window analysis of neuraminidase nucleotide sequences for Australia continent for year 2022.

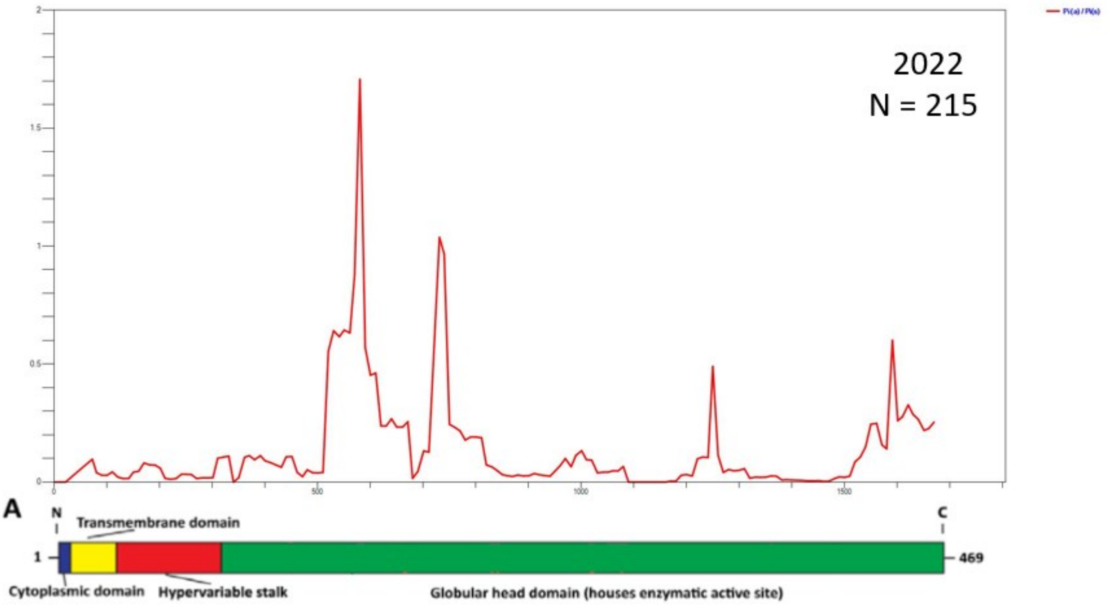
The biggest peak seen on the sliding window analysis of neuraminidase nucleotide sequences for South America is towards the end of the globular head domain, which may change the enzymatic activity of neuraminidase.



Wohldbold et al. 2014

Figure 3.61 Sliding window analysis of neuraminidase nucleotide sequences for South America continent for year 2022.

The sliding window analysis of neuraminidase nucleotide sequences for Africa in 2022 shows that the highest number of nonsynonymous changes happen on the globular head domain of neuraminidase which houses the enzymatic active site. The changes happening on the enzymatic site may enhance the enzymatic activity of neuraminidase, leading to more efficient cleavage of the hemagglutinin-sialic acid bond, which would positively affect how efficient influenza is for spreading through the host body.



Wohldbold et al. 2014

Figure 3.62 Sliding window analysis of neuraminidase nucleotide sequences for Africa continent for year 2022.

Most of the nonsynonymous changes that are seen on the sliding window analyses of neuraminidase nucleotide sequences are on the hypervariable stalk domain and the globular head domain.

The hypervariable stalk domain, as the name suggests, is changing much more and these changes may be affecting the stability of the neuraminidase stalk as well as the height of the stalk. The stability is important for neuraminidase to stay intact. The height of stalk is very important, because of the purpose of the neuraminidase. If the neuraminidase stalk is too short, it will physically prevent the enzymatic site to reach the hemagglutinin-sialic acid links. If the neuraminidase is unable to separate hemagglutinin from the sialic acid receptor, the newly formed viral particles will not be able to leave the host cell and will not be able to spread further. If the neuraminidase stalk is too long, then the enzymatic site will prevent the hemagglutinin from binding to the sialic acid during the initial entrance to the host cell. If the hemagglutinin cannot bind to sialic acid, the viral infection of the host cell cannot be facilitated. The changes on the height of stalk may be positive to the influenza by optimizing the level of enzymatic activity of neuraminidase.

The region between 475-500 nt is observed to be conserved in the sliding window analyses, showing very little to none changes that is nonsynonymous. This region is in the globular head domain that contains the enzymatic active site.

The sliding window analyses for neuraminidase showed that in year 2021, the sequences had less peaks compared to 2022, and even if the biggest peak was not on the hypervariable stalk, there were still some nonsynonymous changes on that region. In 2022, there are more peaks in comparison, which means there are more spots for nonsynonymous changes compared to 2021. However, unlike in 2021, some of the sequences showed very little nonsynonymous changes on the hypervariable stalk region.

CHAPTER IV

CONCLUSION

The hypothesis of this thesis is studying the molecular evolution of 2009 pandemic H1N1 influenza viral sequences can give valuable information about the types of selection acting on the hemagglutinin and neuraminidase proteins of the virus. To test this hypothesis the molecular evolutionary dynamics of 2009pdm-like H1N1 sequences of influenza A were studied with different kinds of phylogenetic and statistical tests.

While testing the hypothesis, it became clear that there is a slight positive selection for all of the HA sequences. The pandemic year of 2009 generally had more negative Tajima's D values, which indicates positive selection or rapid population growth. While the TD values are slightly less negative for intermediate years (2015-2021), the values of 2022 are close to the pandemic years. The 2022 HA nonsynonymous TD values are closer to the pandemic year compared to the other years. Our findings were also supported by the works of Li et al., Furuse et al., de Vries et al., which showed that the 2009 pandemic strain of HA is under positive selection[57 58 59]. The nucleotide diversity of the HA for year 2022 is higher than the pandemic year of 2009, which indicates even though there are more nucleotide diversity, the selection of these mutations is less positive compared to the pandemic year. The year 2021 was a unique year for influenza because there were almost no recorded infections for that year. For HA, this signaled itself as less positive selection on the genomes, without any statistical significance.

The NA sequences showed a very similar trend to the HA sequences as the pandemic year of 2009 and post COVID-19 pandemic year show similar results compared to one another. Once again, the selection on the sequences are positive selections indicated by the Tajima's D and Fu-Li's D* and F* tests. These tests show that there is a positive selection as well as population growth on the sequences. The works of Biggerstaff et al., Correia et al., Xu et al. aligned with our findings, that there is a positive selection acting on the NA sequences[57 60 61]. The population genetics tests with an outgroup showed that there is little to none fixed differences between the human H1N1 2009 pandemic influenza NA and swine H1N1 influenza NA sequences. This suggests a sweeping selection on the NA sequences. As the years go by, the human pandemic H1N1 NA and swine H1H1 NA start to diverge from one another, as there are fixed synonymous and nonsynonymous differences start to appear on the McDonald-Kreitman tests. The work of Xu et al., 2012 also showed that the swine NA sequences are evolutionary close to the human 2009pdm NA sequences and these sequences start to evolutionary diverge from another as the time passes by[61].

Looking at the results, South America region usually has the less negative Tajima's D values, and sometimes these values are not significant. This indicates that the selection on South America region is not as positive compared to the other regions of the world.

With the findings of BEAST analyses and construction of the timetrees for the HA and NA proteins, it is not possible to comment on how the influenza starts spreading around the globe for each season. The findings suggest that there is no certain region that starts the flu season for northern or southern hemisphere. This is further supported by the GISAID phylogenetic timetree analyses for the HA and NA sequences (Supplementary Figure 53, 54). The findings of BEAST analyses also showed that in addition to positive selection on the HA and NA sequences, they align with the exponential growth model. 2009 showed greater exponential growth rates, which is also seen on the work of Biggerstaff et al. where the estimation of R_0 for 2009pdm H1N1 influenza is 1.46[60].

The sliding window analyses of both hemagglutinin and neuraminidase nucleotide sequences showed where on the sequences the most and the least amount of nonsynonymous changes are happening. The sliding window analyses of hemagglutinin showed that most of the changing sites are on the receptor binding domain, and the sites are mostly antigenic sites and receptor binding sites of the molecule. With the positive selection acting on the sequences, the changes on the antigenic sites can be attributed to better immunity evasion of influenza which will help the virus go undetected and spread easier inside the host body. The changes on the receptor binding sites could help the hemagglutinin bind to its receptor more efficiently, which increases the chances of the virus to spread among the host body and among the population. Our findings are in line with the works of de Vries et al., Furuse et al. which found that the antigenic sites and the sites that interact with the HA's receptor are evolving under positive selection, which optimizes HA-sialic acid binding[58 59]. While the HA1 part of hemagglutinin showed many peaks (nonsynonymous changes), the HA2 part of the hemagglutinin showed very few peaks. Looking at the results the HA2 is very conserved.

The sliding window analyses of neuraminidase showed that there are many nonsynonymous changes happening on the globular head domain which contains the enzymatic active site of neuraminidase, as well as the hypervariable stalk of neuraminidase. There were less changes happening on the transmembrane domain of the neuraminidase. Seeing that the neuraminidase sequences are also under positive selection the changes happening on the globular head domain could be attributed to positive changes on the enzymatic activity of the neuraminidase. These changes could be making the neuraminidase enzymatic activity more efficient, which will result in more efficient release of newly formed viral particles from the host cell. The changes on the stalk may be for the optimization of the length and stability of the stalk part of neuraminidase. The findings of the sliding window analyses were in line with the works of Correia et al., Xu et al. which have found that there are positive changes on the head domain of the NA, which contains the enzymatic active site[61 62]. The length of the stalk is very important for proper viral infection cycle. For neuraminidase, the nucleotide positions between 475-500 were conserved, however there were not any other parts of neuraminidase to be found as conserved like in hemagglutinin HA2.

A lot of amino acid changing variants were observed in each year and in each continent. The effects of these polymorphisms on the structure of HA and NA proteins will be studied next in order to examine whether these amino acid variants lead to changes in the protein structure. This information may be valuable for researching novel vaccine or drug (small molecule) targets against H1N1 2009 pandemic like influenza viruses.

REFERENCES

1. Krammer, F., Smith, G. J. D., Fouchier, R. A. M., Peiris, M., Kedzierska, K., Doherty, P. C., Palese, P., Shaw, M. L., Treanor, J., Webster, R. G., & García-Sastre, A. (2018). Influenza. In *Nature Reviews Disease Primers* (Vol. 4, Issue 1, pp. 1–21). Nature Publishing Group. https://doi.org/10.1038/s41572-018-0002-

2. Sautto, G. A., Kirchenbaum, G. A., & Ross, T. M. (2018). Towards a universal influenza vaccine: Different approaches for one goal. In *Virology Journal* (Vol. 15, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12985-017-0918-y

3. "Flu Symptoms & Diagnosis". Centers for Disease Control and Prevention (CDC). 10 July 2019. Retrieved 24 January 2020.

4. Dadonaite, B., Gilbertson, B., Knight, M. L., Trifkovic, S., Rockman, S., Laederach, A., Brown, L. E., Fodor, E., & Bauer, D. L. V. (2019). The structure of the influenza A virus genome. In *Nature Microbiology* (Vol. 4, Issue 11, pp. 1781–1789). Nature Publishing Group. https://doi.org/10.1038/s41564-019-0513-7

5. Wille, M., & Holmes, E. C. (2020). The ecology and evolution of influenza viruses. *Cold Spring Harbor Perspectives in Medicine*, *10*(7), 1–19. https://doi.org/10.1101/cshperspect.a038489

6. Dyason, J. C., & von Itzstein, M. (2010). Viral surface glycoproteins in carbohydrate recognition: structure and modelling. In *Microbial Glycobiology*. https://doi.org/10.1016/B978-0-1237-4546-0.00015-8

7. Liu, S. T. H., Behzadi, M. A., Sun, W., Freyn, A. W., Liu, W. C., Broecker, F., Albrecht, R. A., Bouvier, N. M., Simon, V., Nachbagauer, R., Krammer, F., & Palese, P. (2018). Antigenic sites in influenza H1 hemagglutinin display species-specific immunodominance. *Journal of Clinical Investigation*, *128*(11), 4992–4996. https://doi.org/10.1172/JCI122895

8. Weis, W. I., Brown, J. H., Cusack, S., Paulson, J. C., Skehel, J., & Wiley, D. C. (1988). Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. Nature, 333(6172), 426–431. https://doi.org/10.1038/333426a0

9. Air', G. M., & Laver', W. G. (1989). The Neuraminidase of Influenza Virus. In *PROTEINS: Structure, Function, and Genetics* (Vol. 6).

10. McAuley, J. L., Gilbertson, B. P., Trifkovic, S., Brown, L. E., & McKimm-Breschkin, J. L. (2019). Influenza virus neuraminidase structure and functions. In *Frontiers in Microbiology* (Vol. 10, Issue JAN). Frontiers Media S.A. https://doi.org/10.3389/fmicb.2019.00039

11. Samji, T. (2009). Influenza A: Understanding the Viral Life Cycle. In *YALE JOURNAL OF BIOLOGY AND MEDICINE* (Vol. 82).

12. Coloma, R., Arranz, R., de la Rosa-Trevín, J. M., Sorzano, C. O. S., Munier, S., Carlero, D., Naffakh, N., Ortín, J., & Martín-Benito, J. (2020). Structural insights into influenza A virus ribonucleoproteins reveal a processive helical track as transcription mechanism. *Nature Microbiology*, *5*(5), 727–734. https://doi.org/10.1038/s41564-020-0675-3

13. Zhang, W., Qi, J., Shi, Y., Li, Q., Gao, F., Sun, Y., Lu, X., Lu, Q., Vavricka, C. J., Liu, D., Yan, J., & Gao, G. F. (2010). Crystal structure of the swine-origin A (H1N1)-2009 influenza A virus hemagglutinin (HA) reveals similar antigenicity to that of the 1918 pandemic virus. *Protein and Cell*, *1*(5), 459–467. https://doi.org/10.1007/s13238-010-0059-1

14. Sautto, G. A., & Ross, T. M. (2019). Hemagglutinin consensus-based prophylactic approaches to overcome influenza virus diversity. In *Veterinaria Italiana* (Vol. 55, Issue 3, pp. 195–201). Istituto Zooprofilattico dell'Abruzzo e del Molise. https://doi.org/10.12834/VetIt.1944.10352.1

15. Galloway, S. E., Reed, M. L., Russell, C. J., & Steinhauer, D. A. (2013). Influenza HA Subtypes Demonstrate Divergent Phenotypes for Cleavage Activation and pH of Fusion: Implications for Host Range and Adaptation. *PLoS Pathogens*, *9*(2). https://doi.org/10.1371/journal.ppat.1003151

16. Sriwilaijaroen, N., & Suzuki, Y. (2012). Molecular basis of the structure and function of H1 hemagglutinin of influenza virus. In *Proceedings of the Japan Academy Series B: Physical and Biological Sciences* (Vol. 88, Issue 6, pp. 226–249). https://doi.org/10.2183/pjab.88.226

17. Air', G. M., & Laver', W. G. (1989). The Neuraminidase of Influenza Virus. In *PROTEINS: Structure, Function, and Genetics* (Vol. 6).

18. Varghese, J. N., & Colman, P. M. (1991). Three-dimensional Structure of the Negraminidase of Virus A/Tokyo/3/67 at 2-2 A Resolution. In *I. Mol. Hiol* (Vol. 221).

19. Varghese, J. N., McKimm-Breschkin, J. L., Caldwell, J. B., Kortt, A. A., & Colman CSIRO, P. M. (1992). The Structure of the Complex Between Influenza Virus Neuraminidase and Sialic Acid, the Viral Receptor. In *PROTEINS: Structure, Function, and Genetics* (Vol. 14).

20. Vaccine effectiveness: How well do flu vaccines work? | CDC. (n.d.). https://www.cdc.gov/flu/vaccines-work/vaccineeffect.htm

21. Rajaram, S., Boikos, C., Gelone, D. K., & Gandhi, A. (2020). Influenza vaccines: the potential benefits of cell-culture isolation and manufacturing. In *Therapeutic Advances in Vaccines and Immunotherapy* (Vol. 8). SAGE Publications Ltd. https://doi.org/10.1177/2515135520908121

22. Recommended composition of influenza virus vaccines for use in the 2023-2024 northern hemisphere influenza season. (n.d.). https://www.who.int/publications/m/item/recommended-composition-of-influenza-virus-vaccines-for-use-in-the-2023-2024-northern-hemisphere-influenza-season

23. Fukuyama, H., Shinnakasu, R., & Kurosaki, T. (2020). Influenza vaccination strategies targeting the hemagglutinin stem region. In *Immunological Reviews* (Vol. 296, Issue 1, pp. 132–141). Blackwell Publishing Ltd. https://doi.org/10.1111/imr.12887

24. Jones, S., Nelson-Sathi, S., Wang, Y., Prasad, R., Rayen, S., Nandel, V., Hu, Y., Zhang, W., Nair, R., Dharmaseelan, S., Chirundodh, D. V., Kumar, R., & Pillai, R. M. (2019). Evolutionary, genetic, structural characterization and its functional implications for the influenza A (H1N1) infection outbreak in India from 2009 to 2017. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-51097-w

25. Hegde, N. R. (2015). Cell culture-based influenza vaccines: A necessary and indispensable investment for the future. *Human Vaccines and Immunotherapeutics*, *11*(5), 1223–1234. https://doi.org/10.1080/21645515.2015.1016666

26. Wu, N. C., Zost, S. J., Thompson, A. J., Oyen, D., Nycholat, C. M., McBride, R., Paulson, J. C., Hensley, S. E., & Wilson, I. A. (2017). A structural explanation for the low effectiveness of the seasonal influenza H3N2 vaccine. *PLoS Pathogens*, *13*(10). https://doi.org/10.1371/journal.ppat.1006682

27. Dolgin, E. (2021). mRNA flu shots move into trials. In *Nature reviews. Drug discovery* (Vol. 20, Issue 11, pp. 801–803). NLM (Medline). https://doi.org/10.1038/d41573-021-00176-7

28. Du, X., Dong, L., Lan, Y., Peng, Y., Wu, A., Zhang, Y., Huang, W., Wang, D., Wang, M., Guo, Y., Shu, Y., & Jiang, T. (2012). Mapping of H3N2 influenza antigenic evolution in China reveals a strategy for vaccine strain recommendation. *Nature Communications*, *3*. https://doi.org/10.1038/ncomms1710

29. Jain, S., Finelli, L., Shaw, M., Lindstrom, S., Gubareva, L. V., Xu, X., & Uyeki, T. M. (2009). Emergence of a Novel Swine-Origin Influenza A (H1N1) Virus in Humans. The New England Journal of Medicine, 360(25), 2605–2615. https://doi.org/10.1056/nejmoa0903810

30. Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C. A., Smith, D. J., Pybus, O. G., Brockmann, D., & Suchard, M. A. (2014). Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. *PLoS Pathogens*, *10*(2). https://doi.org/10.1371/journal.ppat.1003932

31. Tscherne, D. M., & García-Sastre, A. (2011). Virulence determinants of pandemic influenza viruses. *Journal of Clinical Investigation*, *121*(1), 6–13. https://doi.org/10.1172/JCI44947

32. Geneva. (1999). *Influenza Pandemic Plan. The Role of WHO and Guidelines for National and Regional Planning*. http://www.who.int/emc

33. Michaelis, M., Doerr, H. W., & Cinatl, J. (2009). Novel swine-origin influenza A virus in humans: Another pandemic knocking at the door. In *Medical Microbiology and Immunology* (Vol. 198, Issue 3, pp. 175–183). https://doi.org/10.1007/s00430-009-0118-5

34. Centers for Disease Control Prevention (CDC) (24 April 2009). "Swine Influenza A (H1N1) Infection in Two Children – Southern California, March–April 2009". Morbidity and Mortality Weekly Report. CDC. 58 (15): 400–402. PMID 19390508.

35. Trifonov, V., Khiabanian, H., & Rabadan, R. (2009). *Geographic Dependence, Surveillance, and Origins of the 2009 Influenza A (H1N1) Virus*. www.who.int/csr/

36. Khanna, M., Saxena, L., Gupta, A., Kumar, B., & Rajput, R. (2013). Influenza pandemics of 1918 and 2009: A comparative account. In *Future Virology* (Vol. 8, Issue 4, pp. 335–342). https://doi.org/10.2217/fvl.13.18

37. Kelly, H., Peck, H. A., Laurie, K. L., Wu, P., Nishiura, H., & Cowling, B. J. (2011). The age-specific cumulative incidence of infection with pandemic influenza H1N1 2009 was similar in various countries prior to vaccination. In *PLoS ONE* (Vol. 6, Issue 8). Public Library of Science. https://doi.org/10.1371/journal.pone.0021828

38. Nelson, M. I., Wentworth, D. E., Culhane, M. R., Vincent, A. L., Viboud, C., LaPointe, M. P., Lin, X., Holmes, E. C., & Detmer, S. E. (2014). Introductions and Evolution of Human-Origin Seasonal Influenza A Viruses in Multinational Swine Populations. *Journal of Virology*, *88*(17), 10110–10119. https://doi.org/10.1128/jvi.01080-14

39. Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. In *Eurosurveillance* (Vol. 22, Issue 13). European Centre for Disease Prevention and Control (ECDC). https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

40. Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sanchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution*, *34*(12), 3299–3302. https://doi.org/10.1093/molbev/msx248

41. Katoh, K., Rozewicki, J., & Yamada, K. D. (2018). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *20*(4), 1160–1166. https://doi.org/10.1093/bib/bbx108

42. https://www.biostars.org/p/183279/ (Retrieved January 2023)

43. Stoletzki, N., & Eyre-Walker, A. (2011). Estimation of the neutrality index. *Molecular Biology and Evolution*, *28*(1), 63–70. https://doi.org/10.1093/molbev/msq249

44. Nei, M., & Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases (molecular evolution/mitochondrial DNA/nucleotide diversity). In *Genetics* (Vol. 76, Issue 10).

45. Watterson, G. A. (1975). On the Number of Segregating Sites in Genetical Models without Recombination. In *POPULATION BIOLOGY* (Vol. 7).

46. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 123(3), 585–595. https://doi.org/10.1093/genetics/123.3.585

47. Fu, Y., & Li, W. (1993). Statistical tests of neutrality of mutations. Genetics, 133(3), 693–709. https://doi.org/10.1093/genetics/133.3.693

48. Fay, J. C., & Wu, C. (2000). Hitchhiking under positive Darwinian selection. Genetics, 155(3), 1405–1413. https://doi.org/10.1093/genetics/155.3.1405

49. Fay, J. C., Wyckoff, G. J., & Wu, C. (2002). Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature, 415(6875), 1024–1026. https://doi.org/10.1038/4151024a

50. McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature, 351(6328), 652–654. https://doi.org/10.1038/351652a0

51. Rand, D. M., & Kann, L. (1996). Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. Molecular Biology and Evolution, 13(6), 735–748. https://doi.org/10.1093/oxfordjournals.molbev.a025634

52. Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, *4*(1). https://doi.org/10.1093/ve/vey016

53. Baele, G., Li, W. L. S., Drummond, A. J., Suchard, M. A., & Lemey, P. (2013). Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Molecular Biology and Evolution*, *30*(2), 239–243. https://doi.org/10.1093/molbev/mss243

54. Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904. https://doi.org/10.1093/sysbio/syy032

55. http://tree.bio.ed.ac.uk/software/figtree/

56. Bianchini, G., & Sánchez-Baracaldo, P. (2023). TreeViewer - Cross-platform software to draw phylogenetic trees (Version 2.1.0) [Computer software]. https://doi.org/10.5281/zenodo.7768344

57. Li, W., Shi, W., Qiao, H., Ho, S. Y., Luo, A., Zhang, Y., &amp; Zhu, C. (2011). Positive selection on hemagglutinin and neuraminidase genes of H1N1 influenza viruses. Virology Journal, 8(1). https://doi.org/10.1186/1743-422x-8-183

58. Furuse, Y., Shimabukuro, K., Odagiri, T., Sawayama, R., Okada, T., Khandaker, I., Suzuki, A., & Oshitani, H. (2010). Comparison of selection pressures on the HA gene of pandemic (2009) and seasonal human and swine influenza A H1 subtype viruses. Virology, 405(2), 314–321. https://doi.org/10.1016/j.virol.2010.06.018

59. De Vries, R. P., De Vries, E., Martínez-Romero, C., McBride, R., Van Kuppeveld, F. J. M., Rottier, P. J. M., García-Sastre, A., Paulson, J. C., & De Haan, C. a. M. (2013). Evolution of the hemagglutinin protein of the new pandemic H1N1 influenza virus: Maintaining optimal receptor binding by compensatory substitutions. Journal of Virology, 87(24), 13868–13877. https://doi.org/10.1128/jvi.01955-13

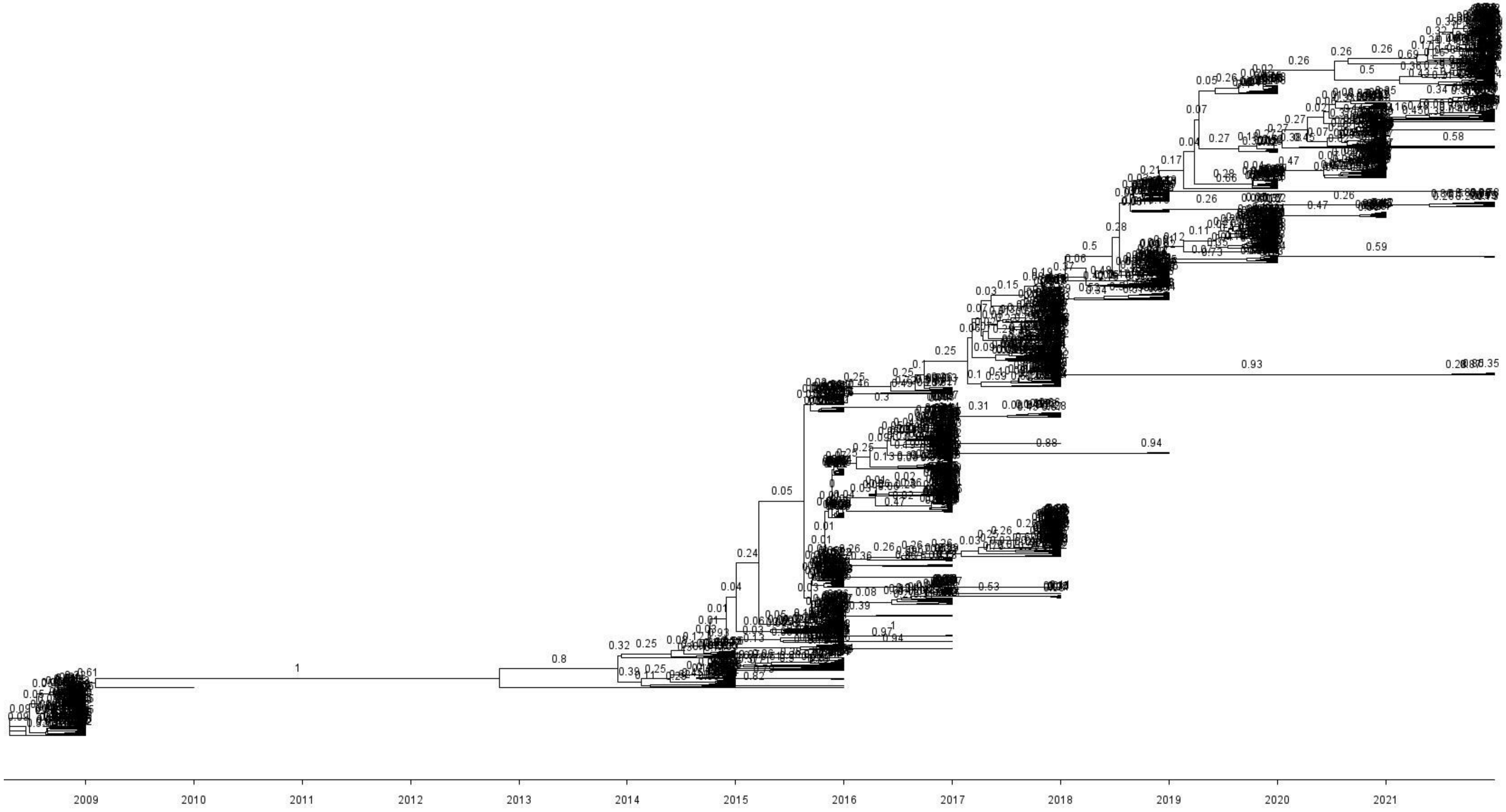
60. Biggerstaff, M., Cauchemez, S., Reed, C., Gambhir, M., & Finelli, L. (2014). Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. BMC Infectious Diseases, 14(1). https://doi.org/10.1186/1471-2334-14-480

61. Xu, J., Davis, C. T., Christman, M. C., Rivailler, P., Zhong, H. A., Donis, R. O., & Lu, G. (2012). Evolutionary History and Phylodynamics of Influenza A and B Neuraminidase (NA) Genes Inferred from Large-Scale Sequence Analyses. PLOS ONE, 7(7), e38665. https://doi.org/10.1371/journal.pone.0038665

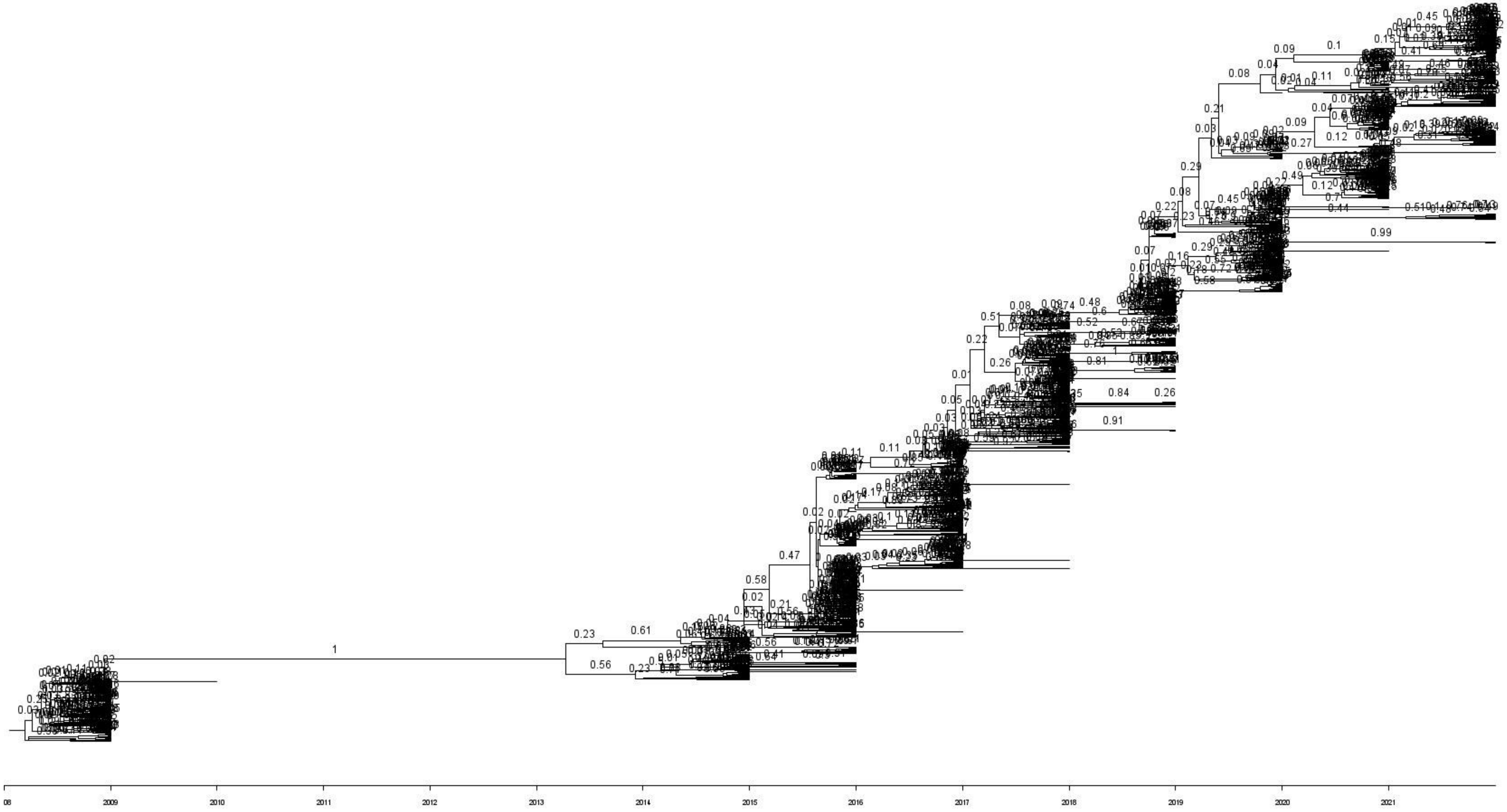
62. Correia, V., Abecasis, A., & Rebelo-De-Andrade, H. (2018). Molecular footprints of selective pressure in the neuraminidase gene of currently circulating human influenza subtypes and lineages. Virology, 522, 122–130. https://doi.org/10.1016/j.virol.2018.07.002

APPENDICES

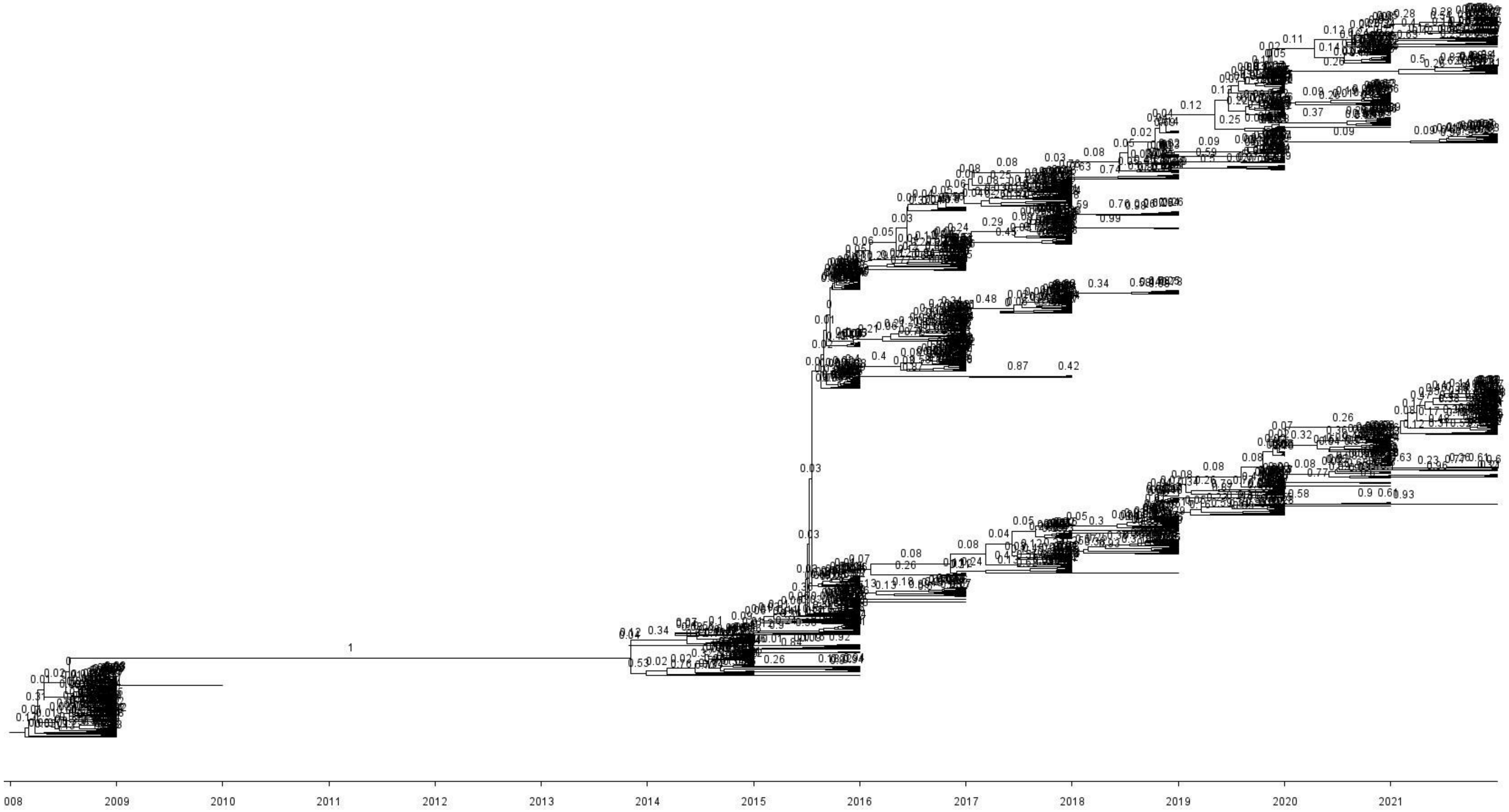
APPENDIX A



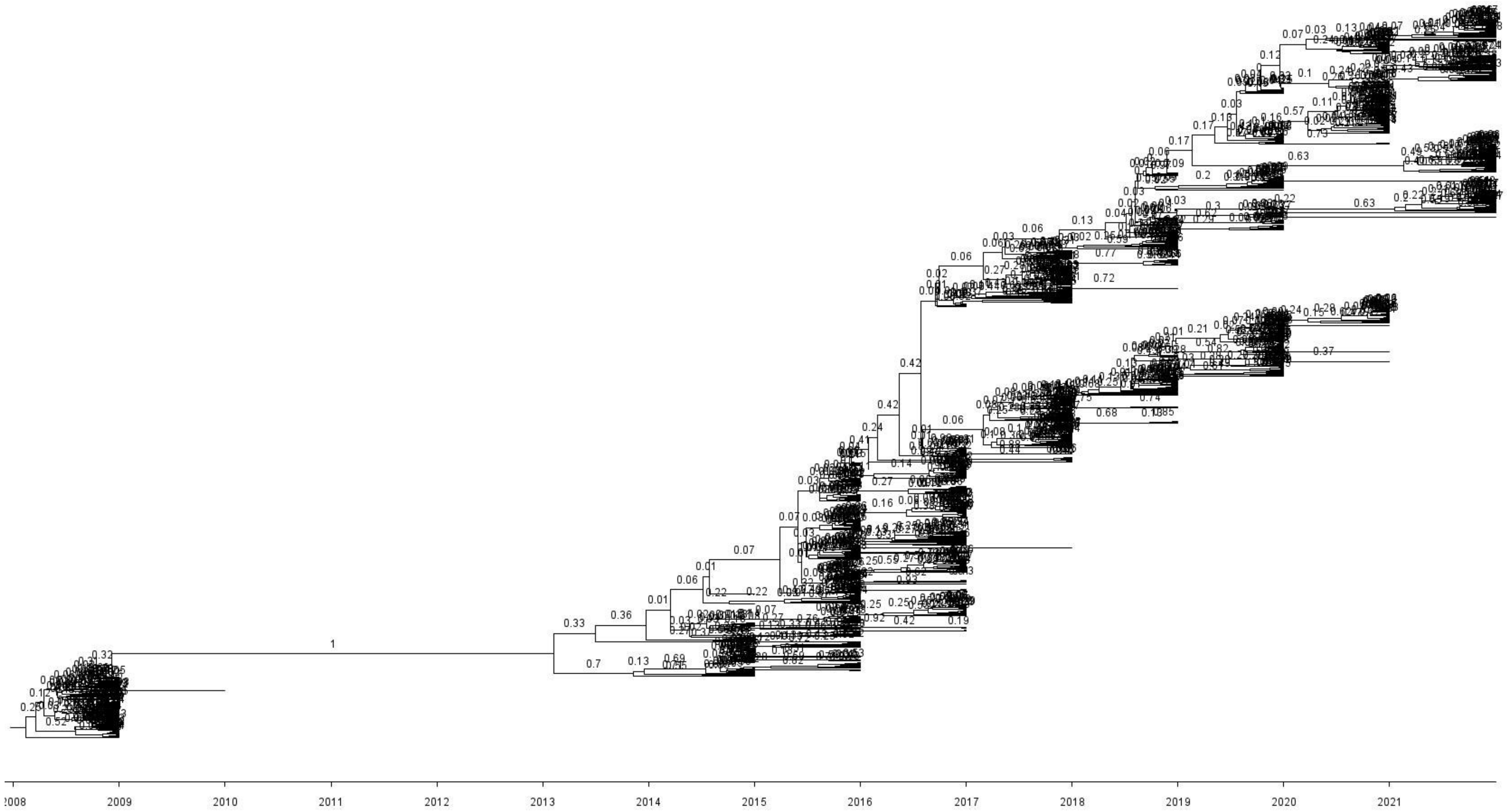
Supplementary Figure 1. Africa consensus HA timetree with posterior probabilities



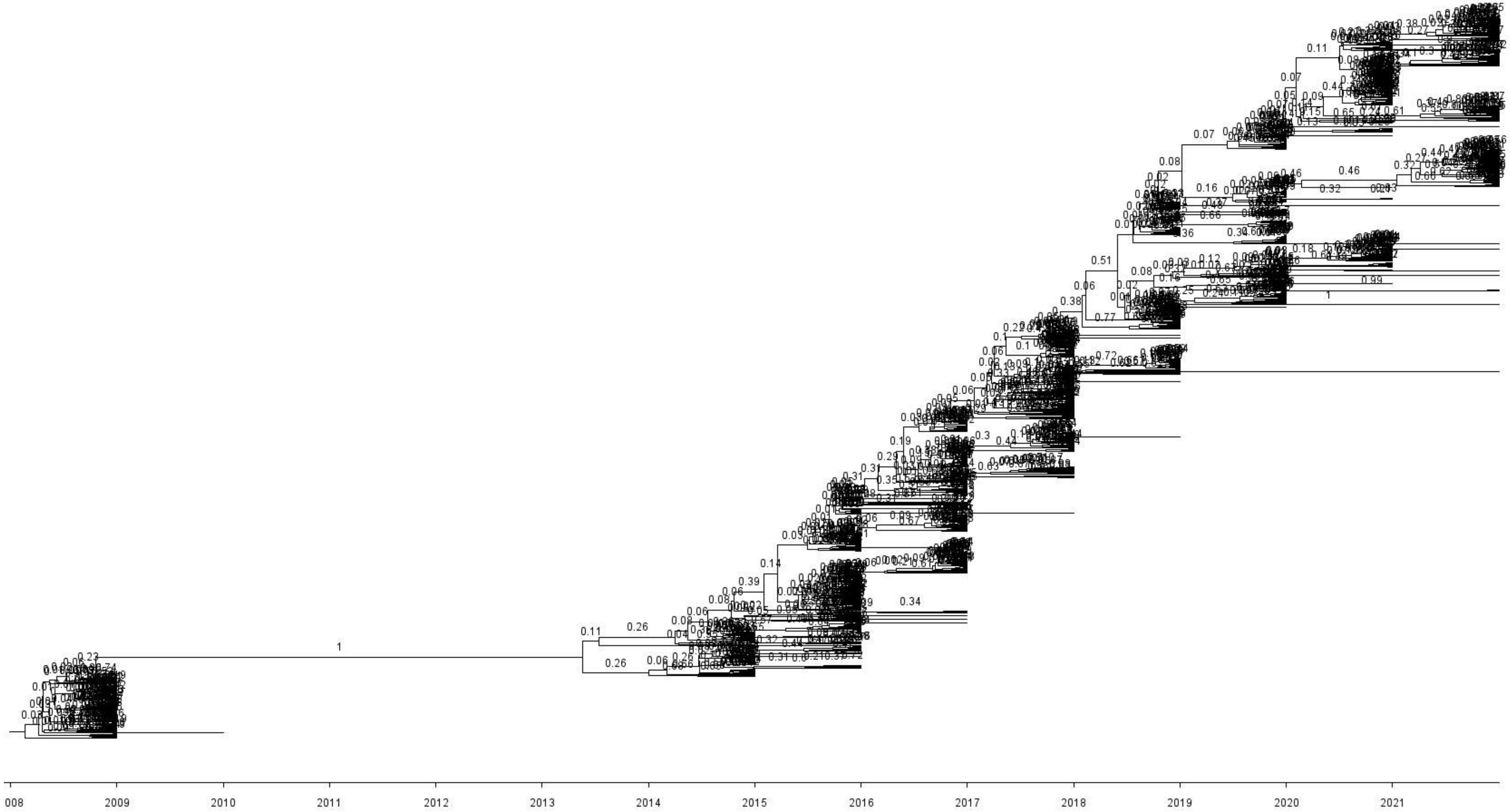
Supplementary Figure 2. Africa HA timetree 1 with posterior probabilities



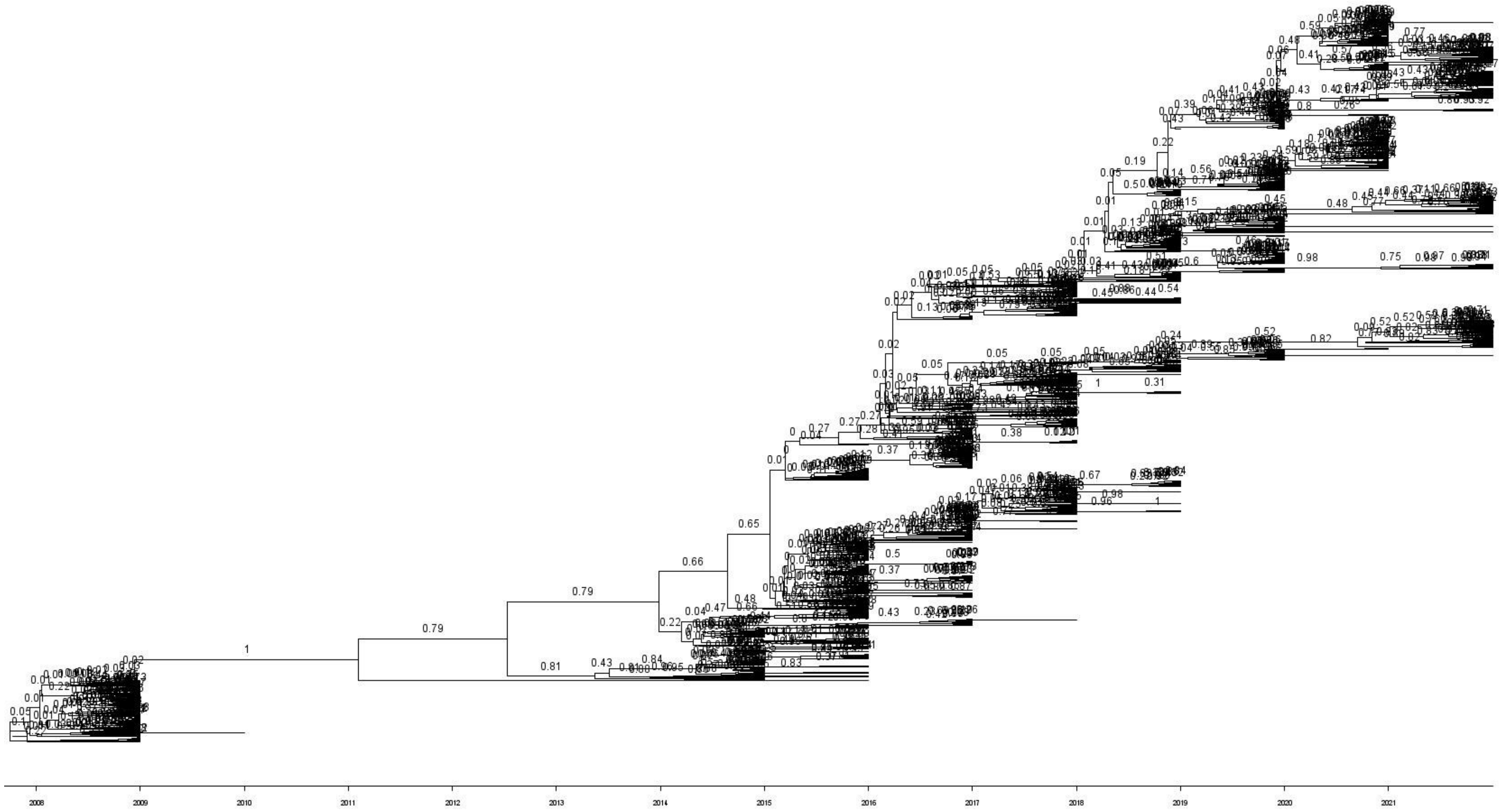
Supplementary Figure 3. Africa HA timetree 2 with posterior probabilities



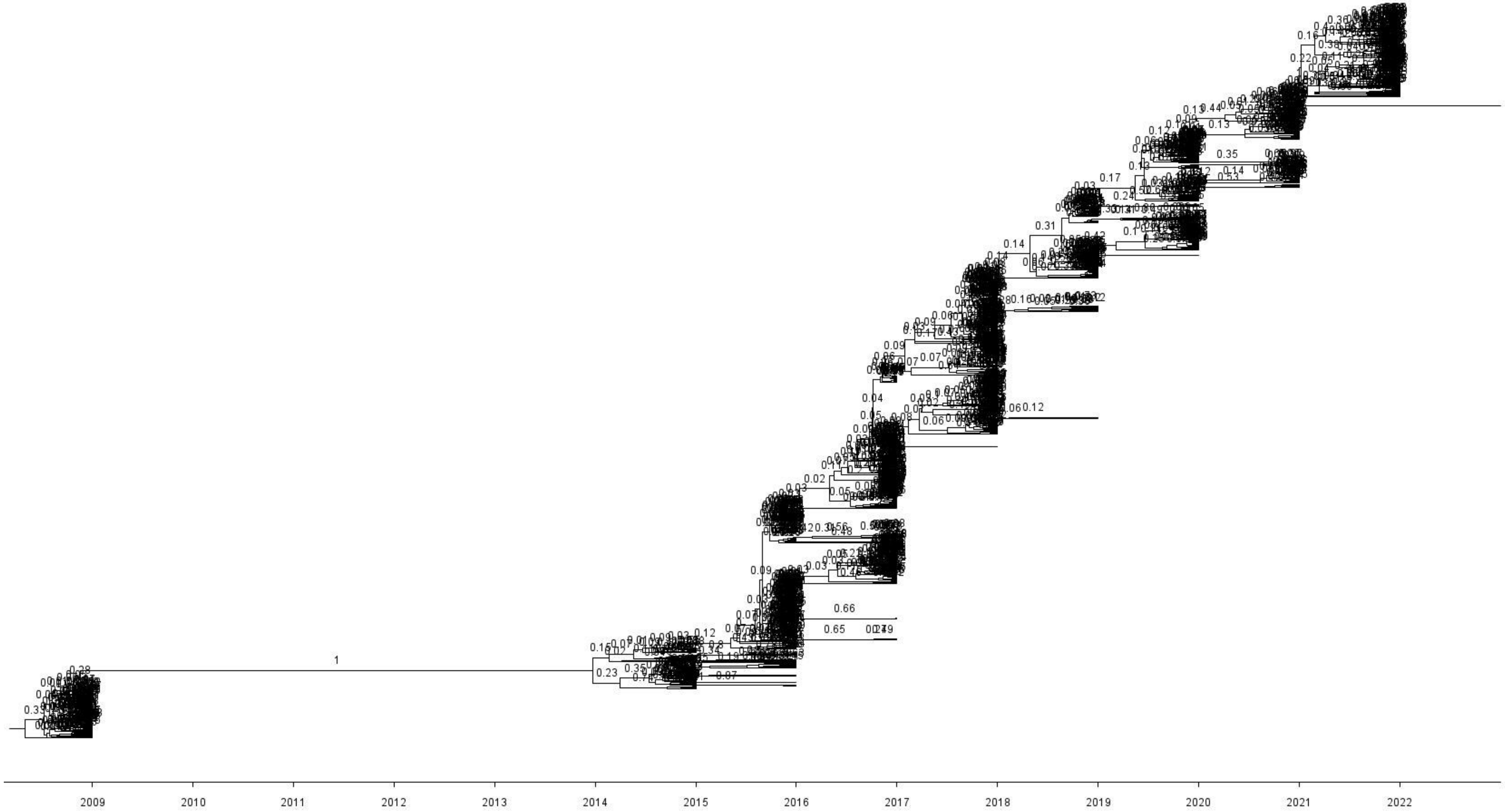
Supplementary Figure 4. Africa HA timetree 3 with posterior probabilities



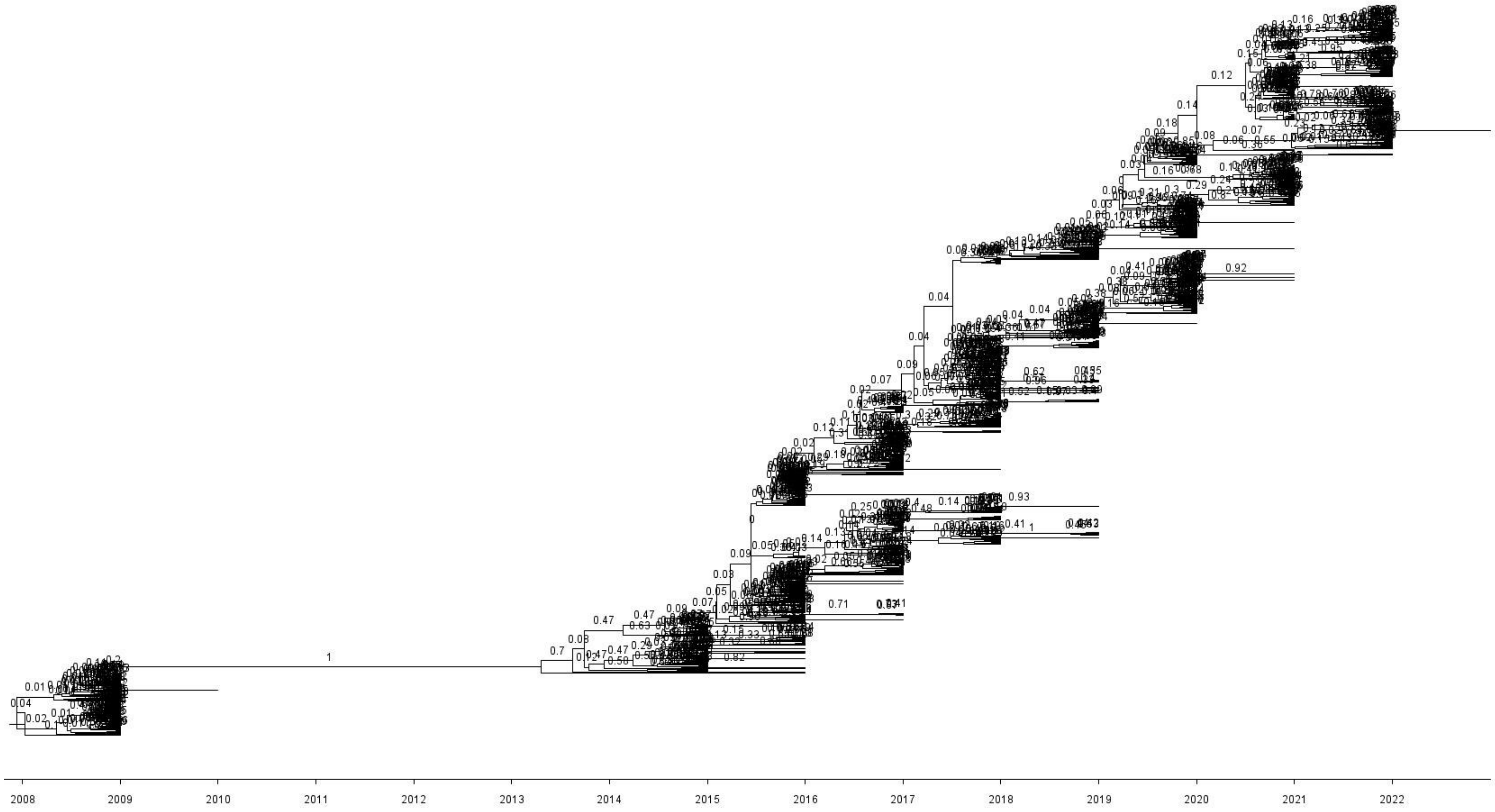
Supplementary Figure 4. Africa HA timetree 3 with posterior probabilities



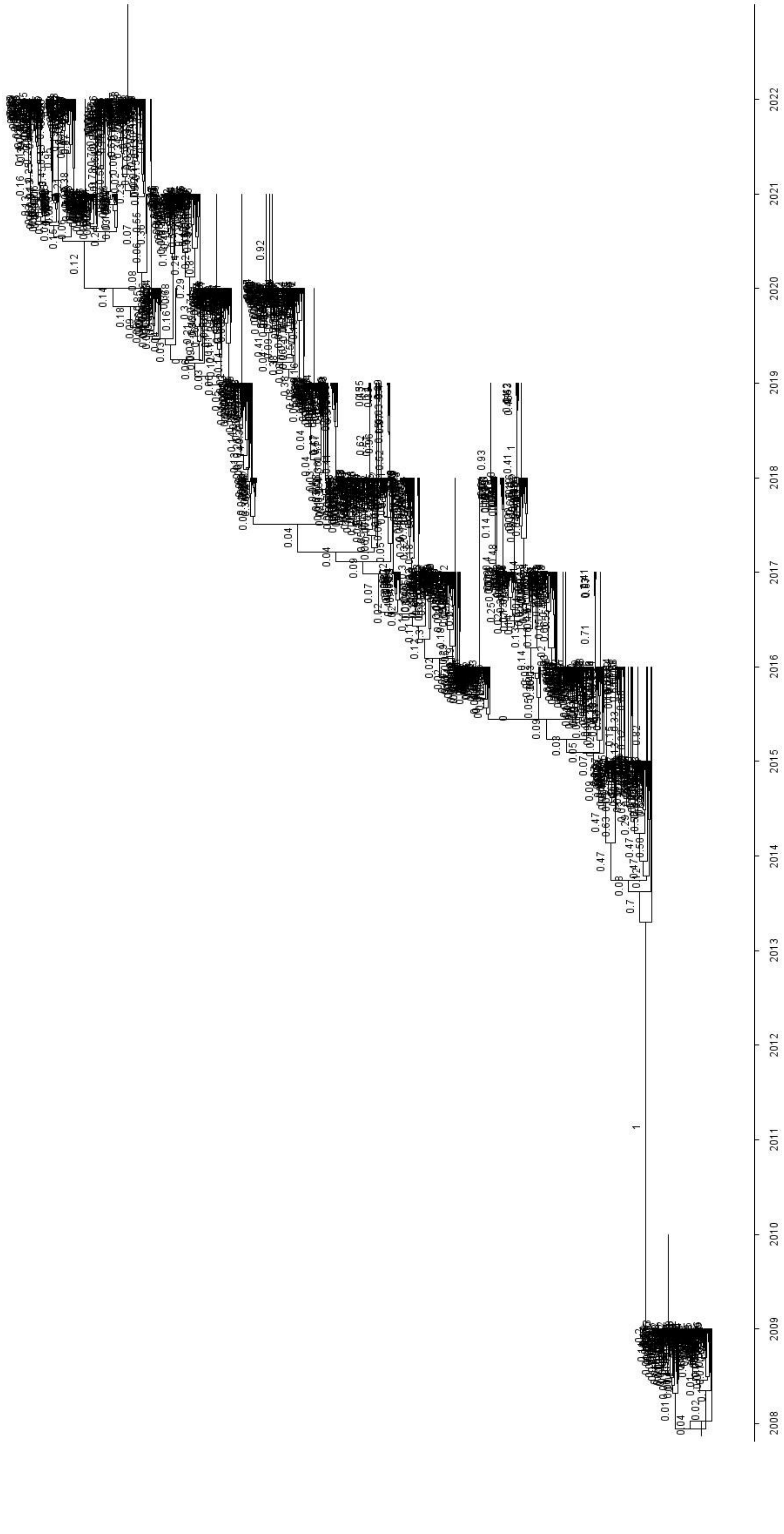
Supplementary Figure 5. Africa HA timetree 4 with posterior probabilities



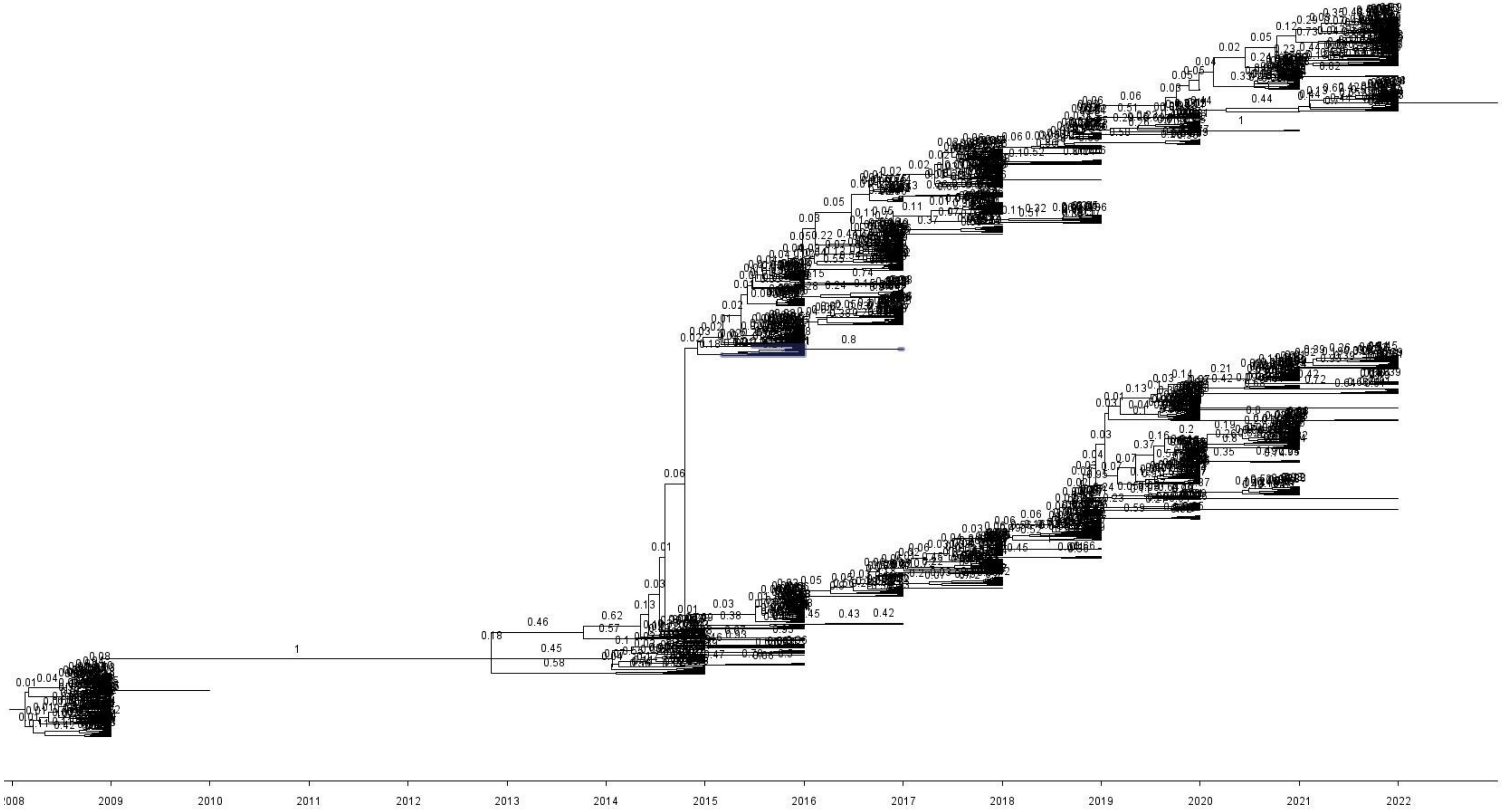
Supplementary Figure 6. Africa NA consensus timetree with posterior probabilities



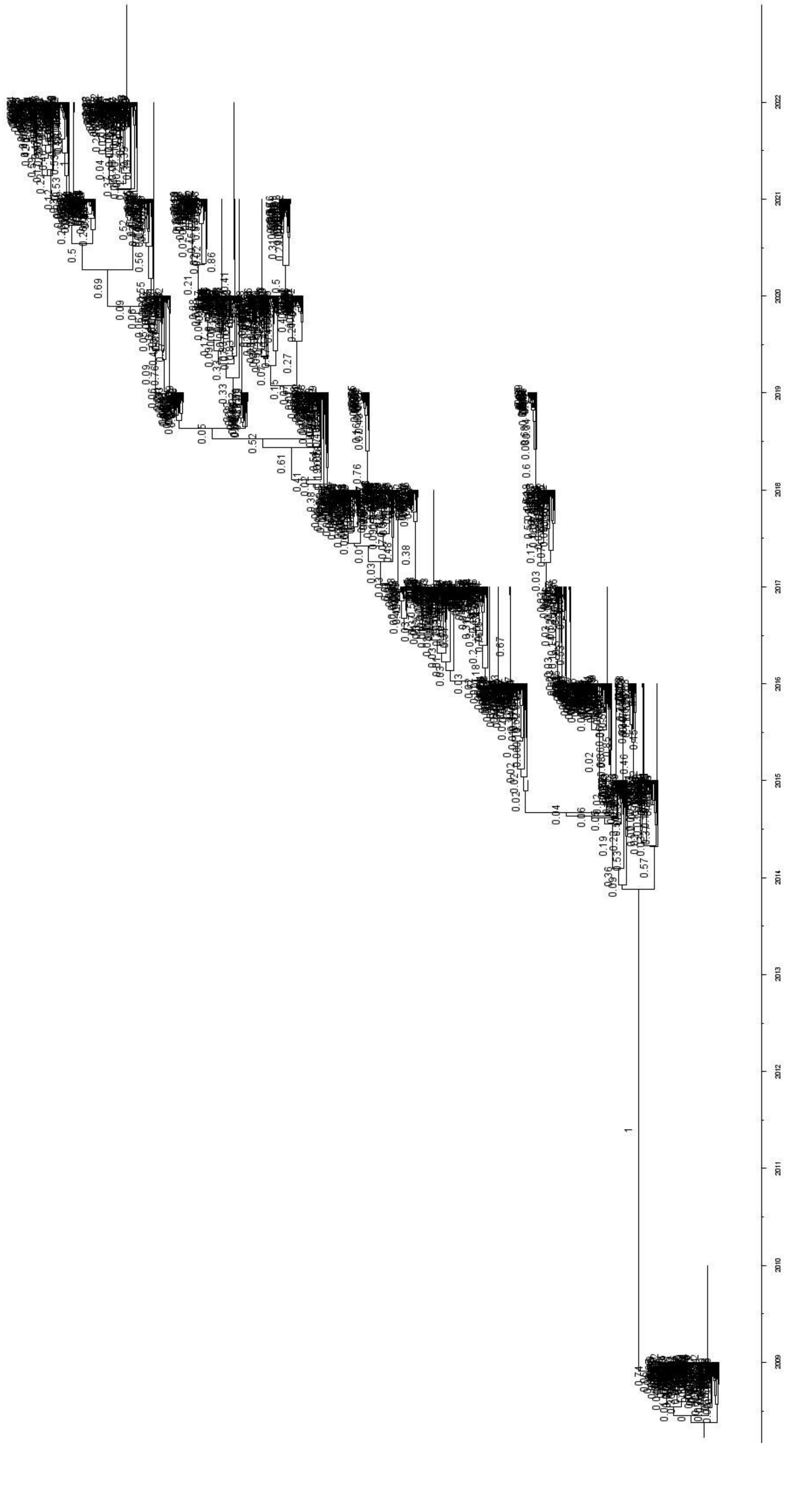
Supplementary Figure 7. Africa NA timetree 1 with posterior probabilities



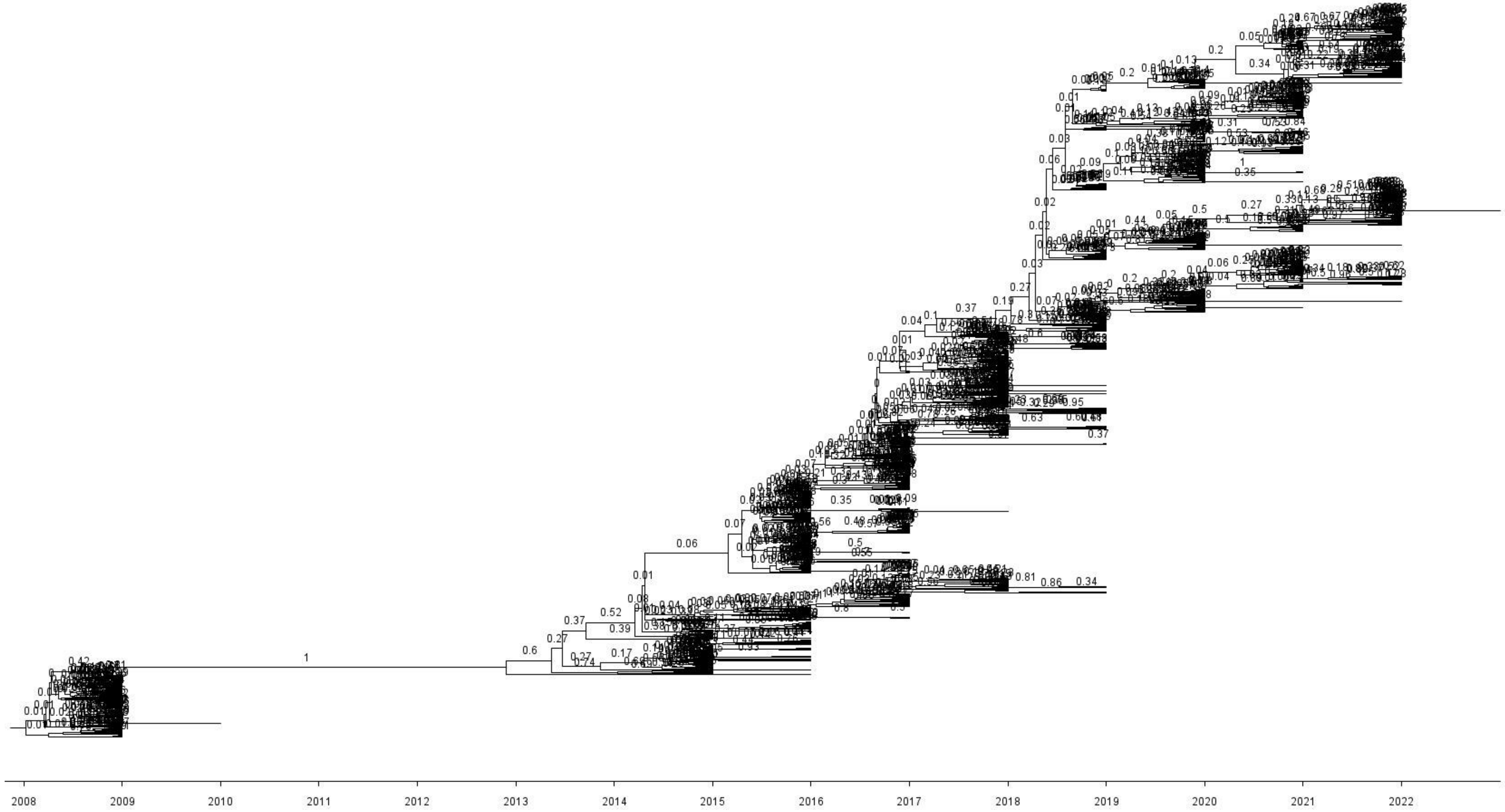
Supplementary Figure 8. Africa NA timetree 2 with posterior probabilities



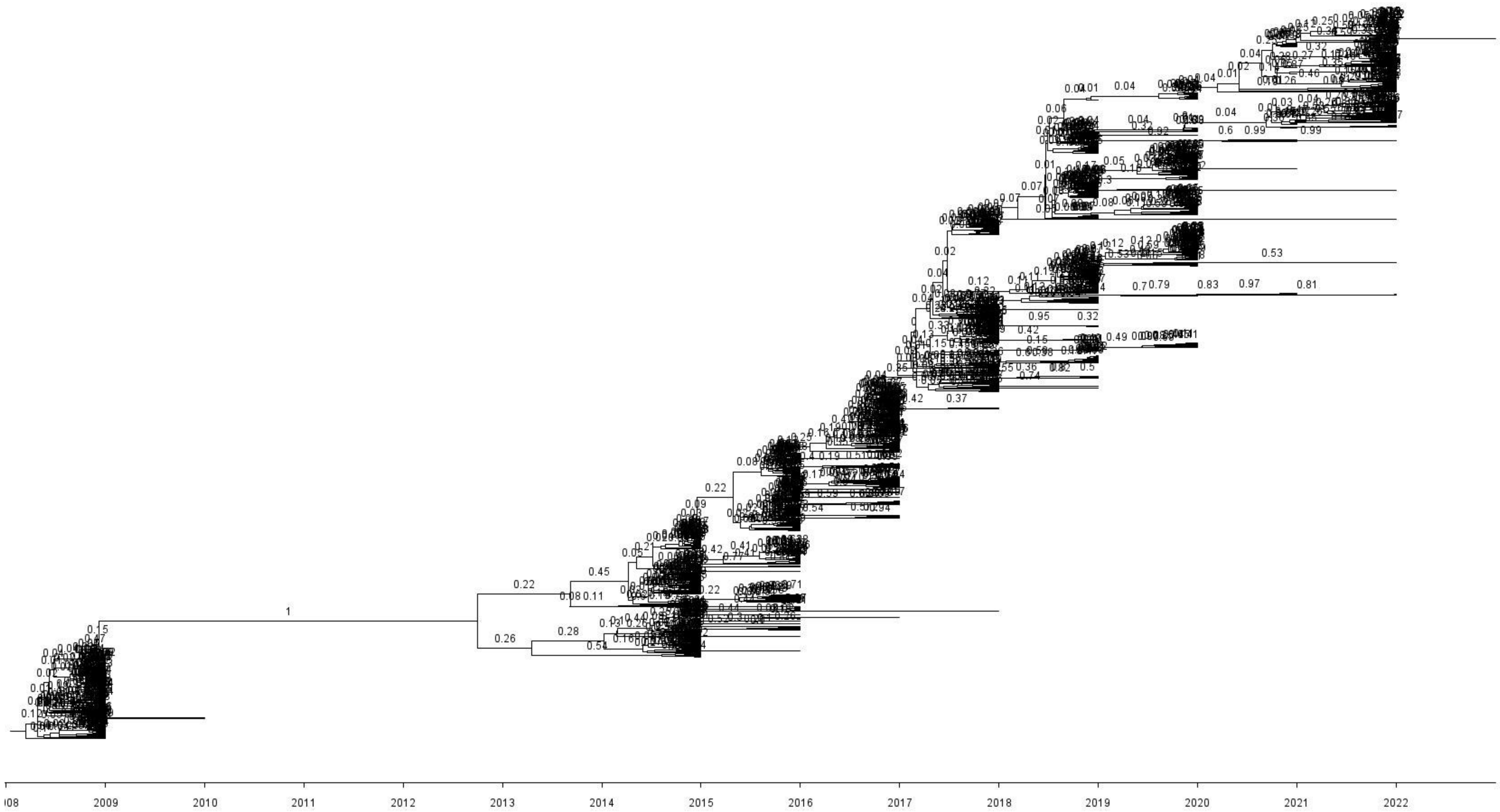
Supplementary Figure 10. Africa NA timetree 4 with posterior probabilities



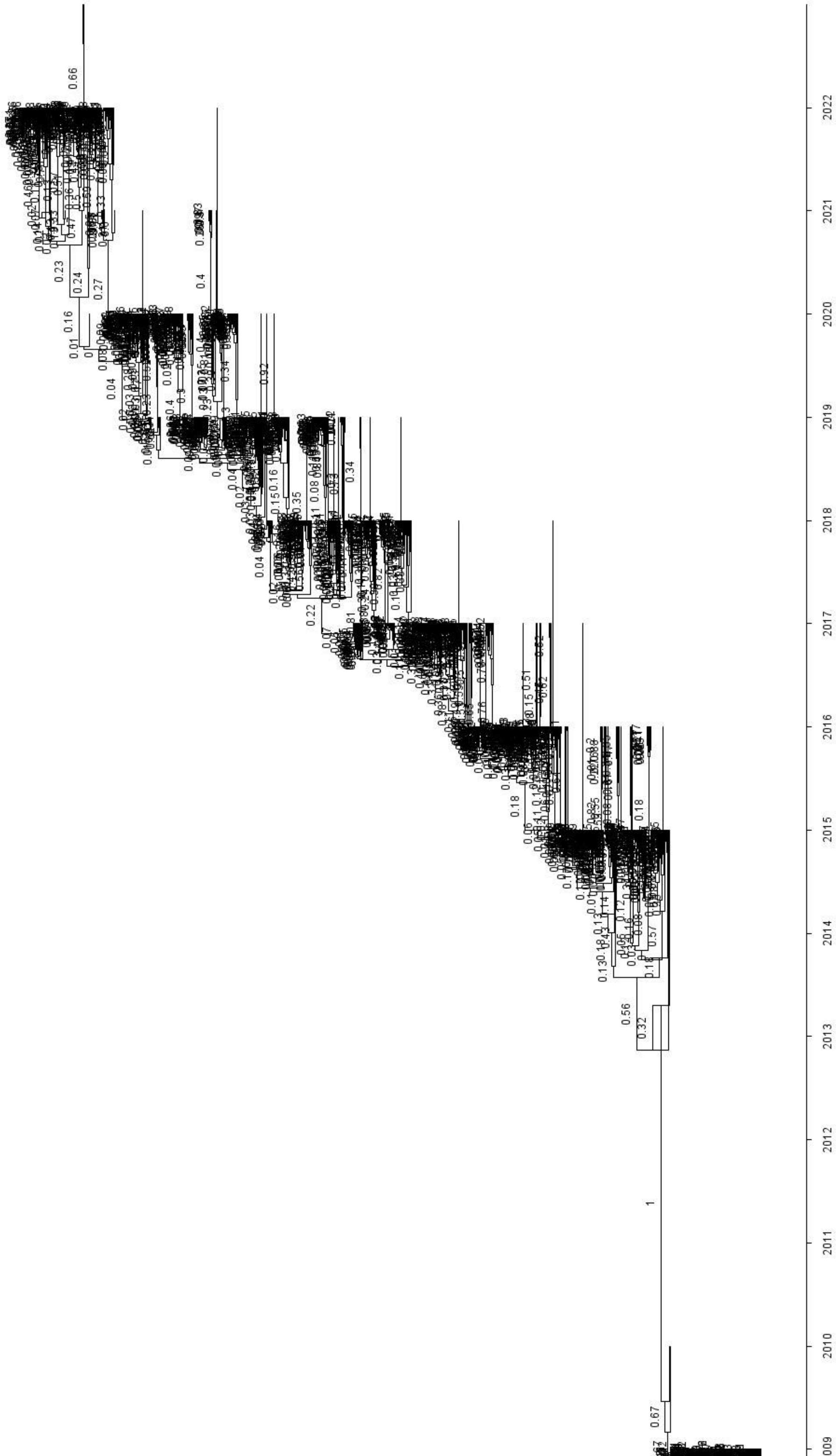
Supplementary Figure 10. Africa NA timetree 4 with posterior probabilities



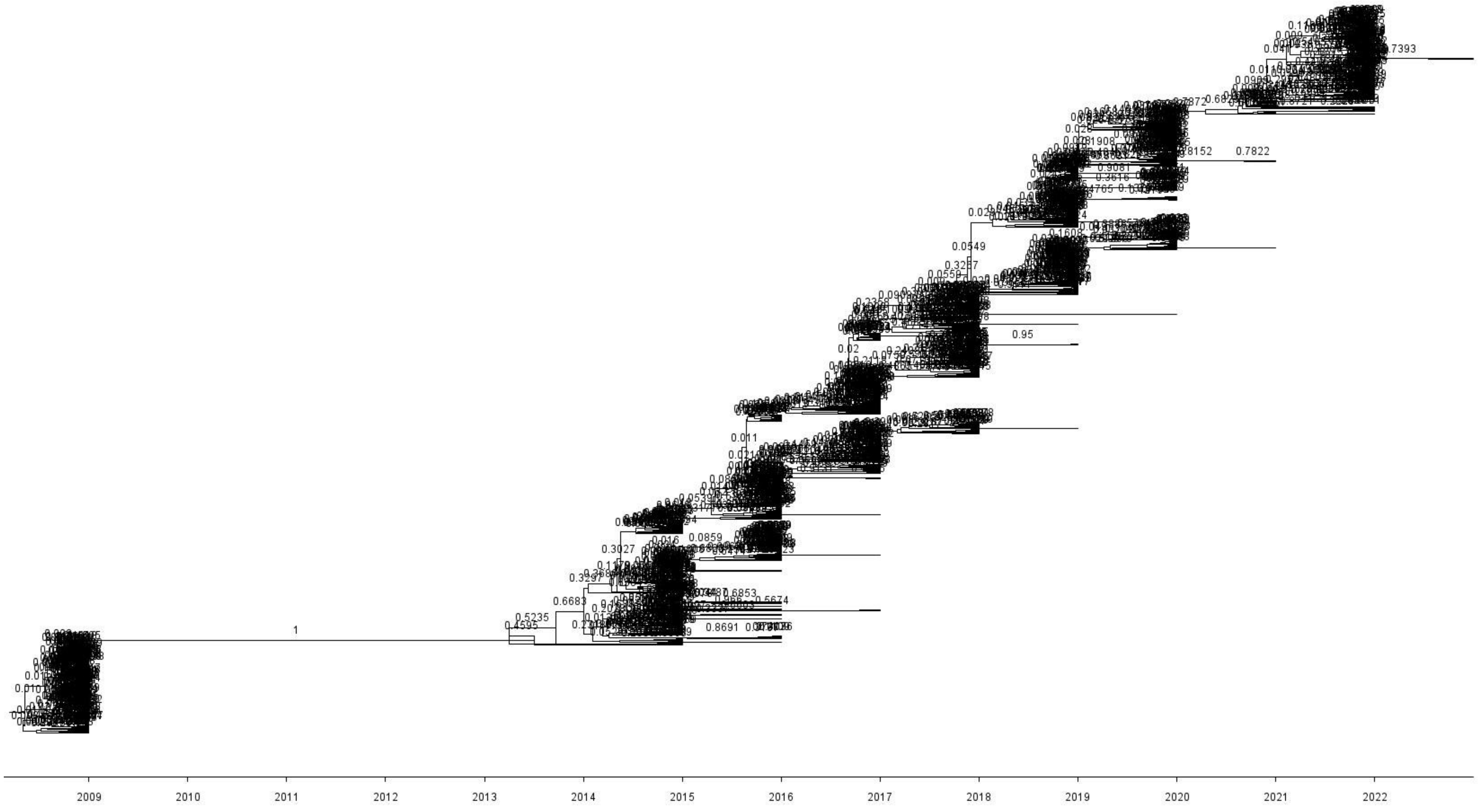
Supplementary Figure 11. Africa NA timetree 5 with posterior probabilities



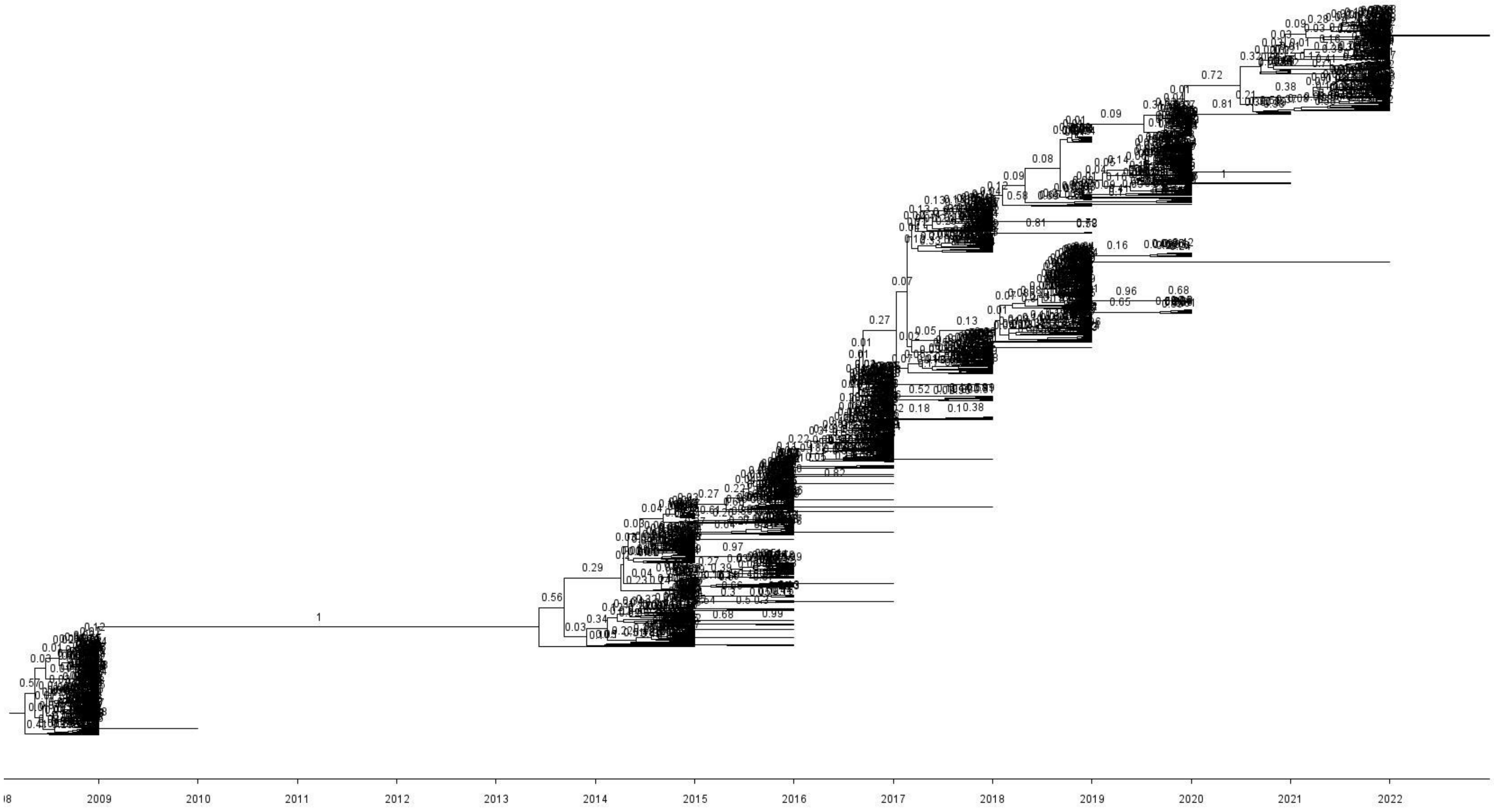
Supplementary Figure 12. Asia HA timetree 1 with posterior probabilities



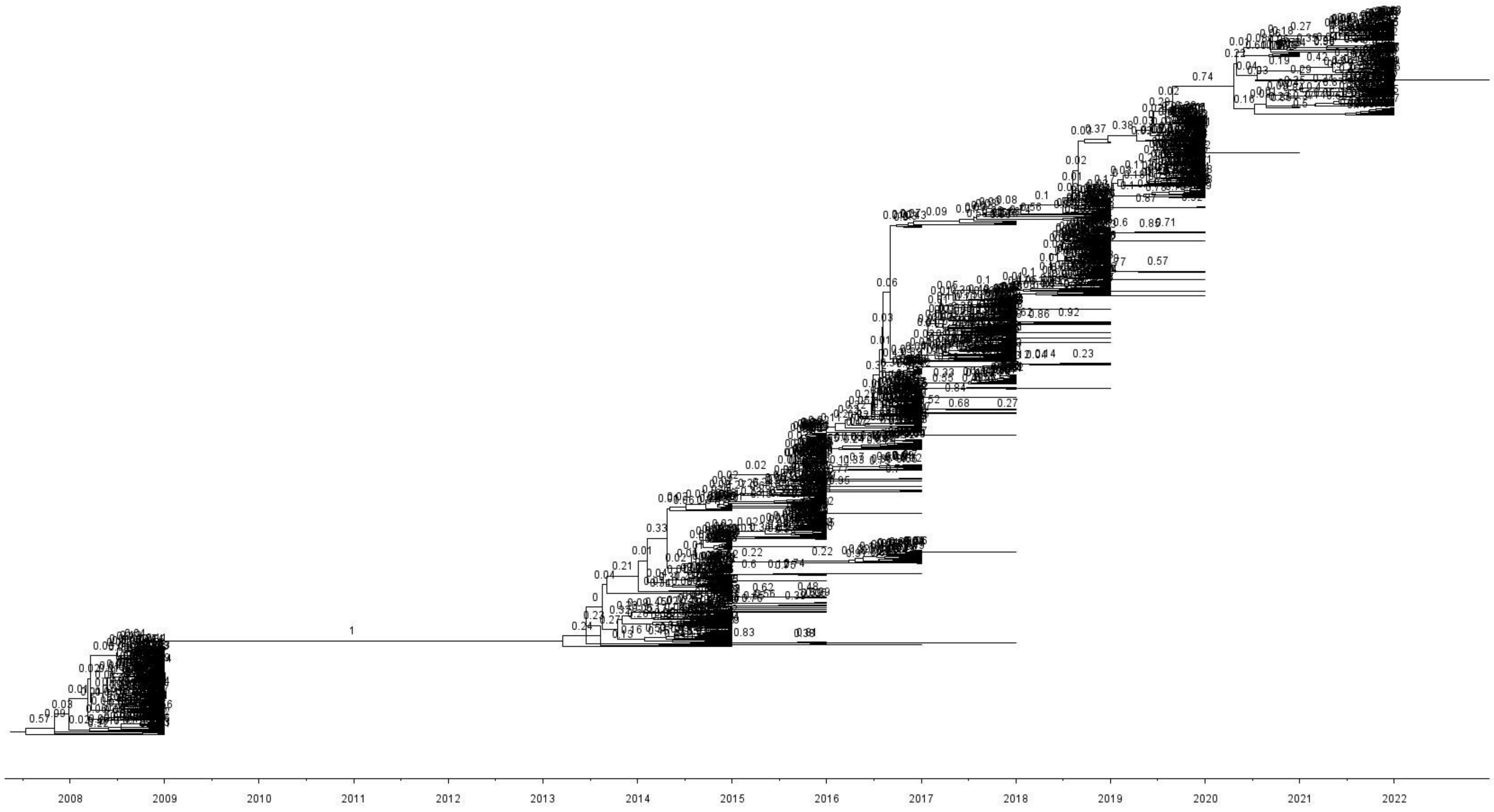
Supplementary Figure 13. Asia HA timetree 2 with posterior probabilities



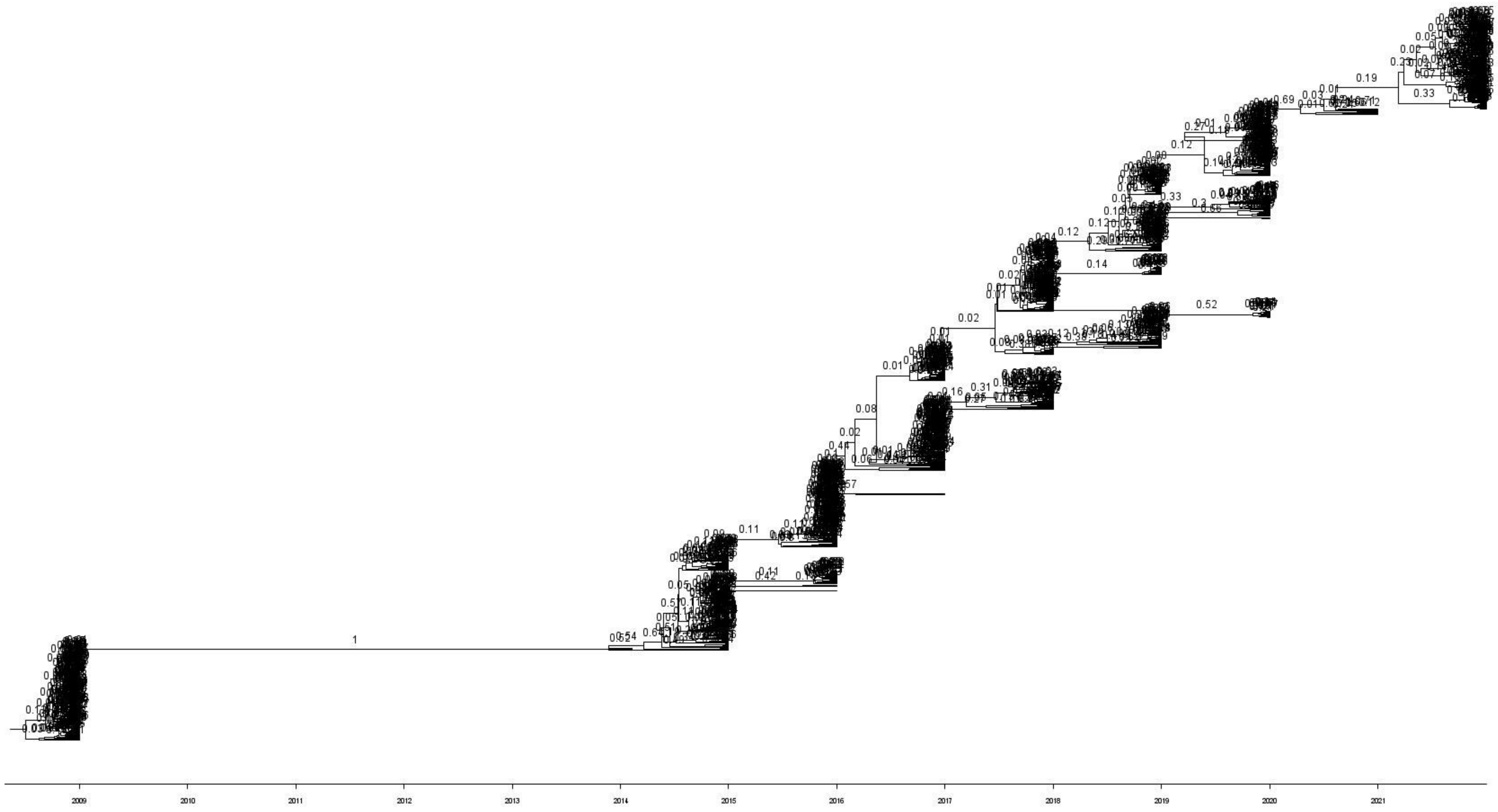
Supplementary Figure 14. Asia HA timetree 3 with posterior probabilities



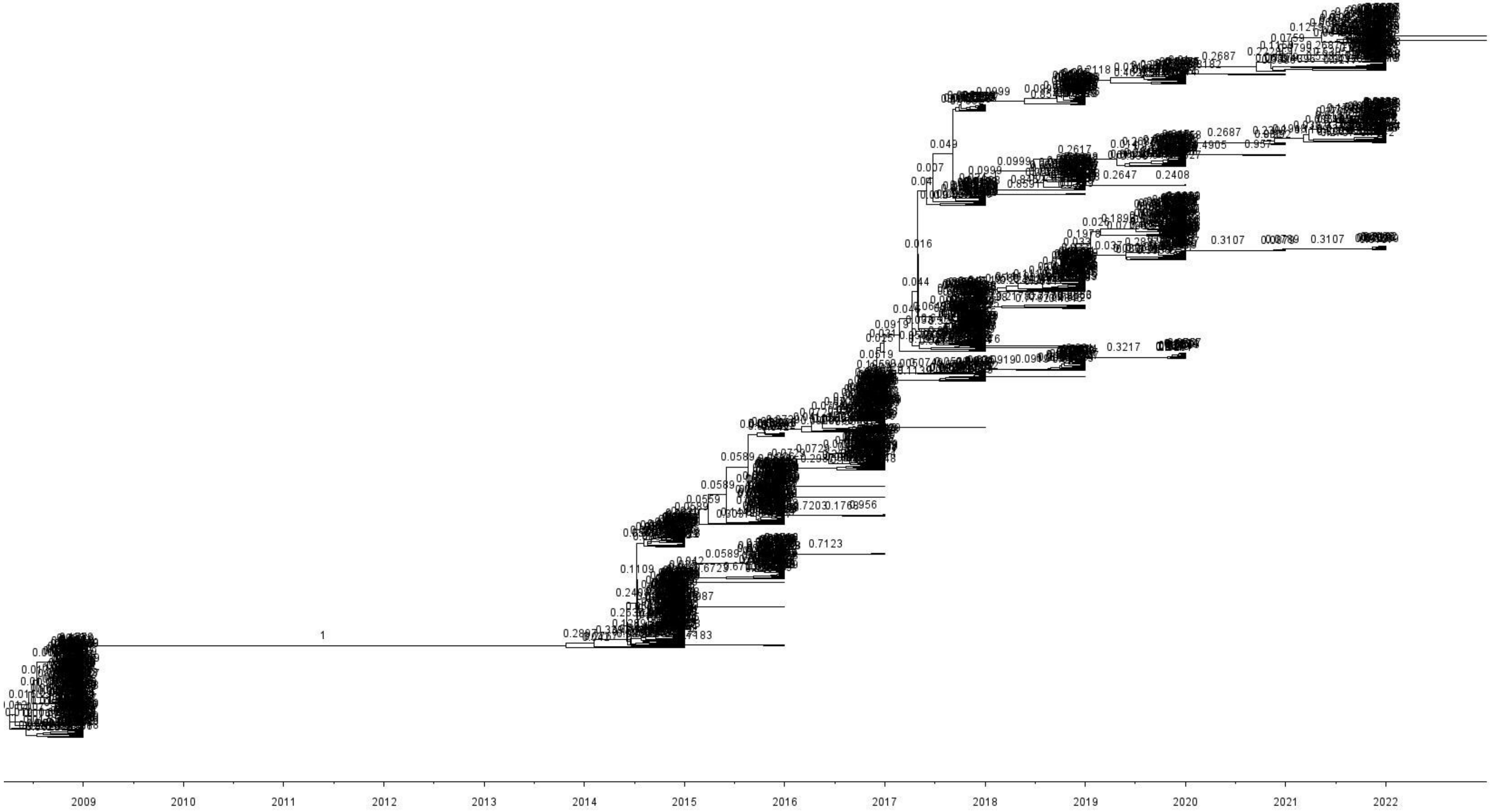
Supplementary Figure 15. Asia HA timetree 4 with posterior probabilities



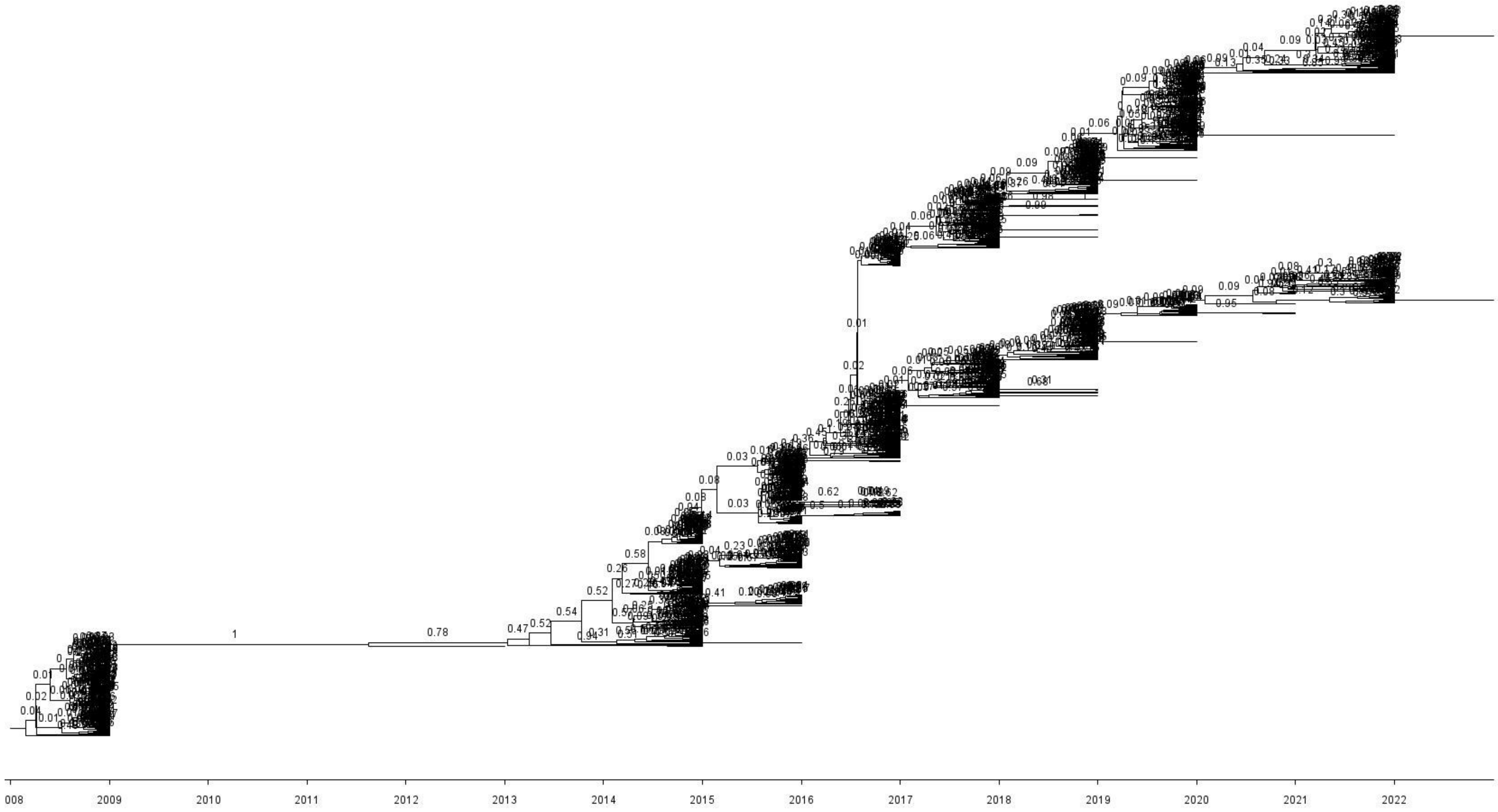
Supplementary Figure 16. Asia HA timetree 5 with posterior probabilities



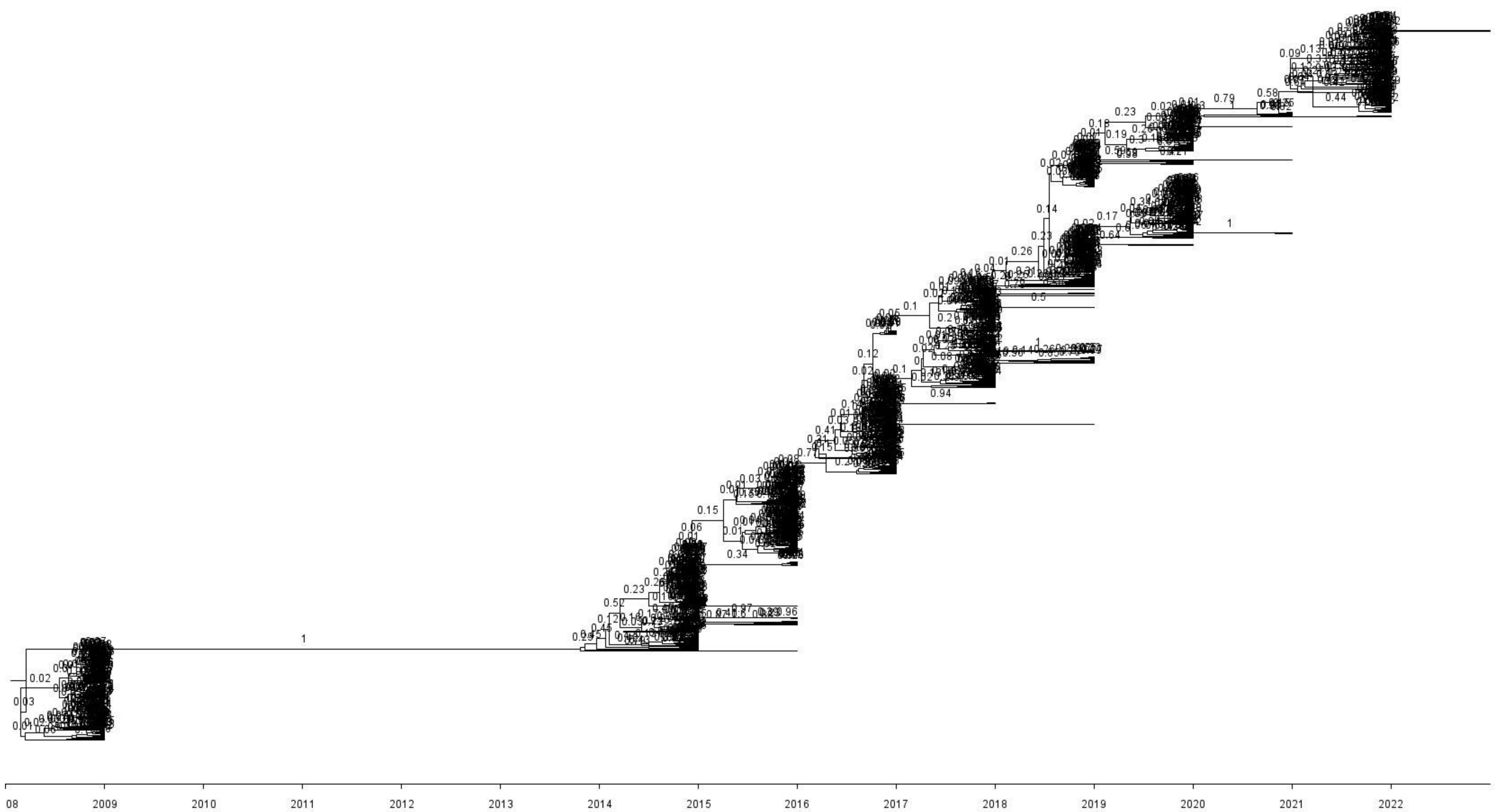
Supplementary Figure 17. Asia NA timetree 1 with posterior probabilities



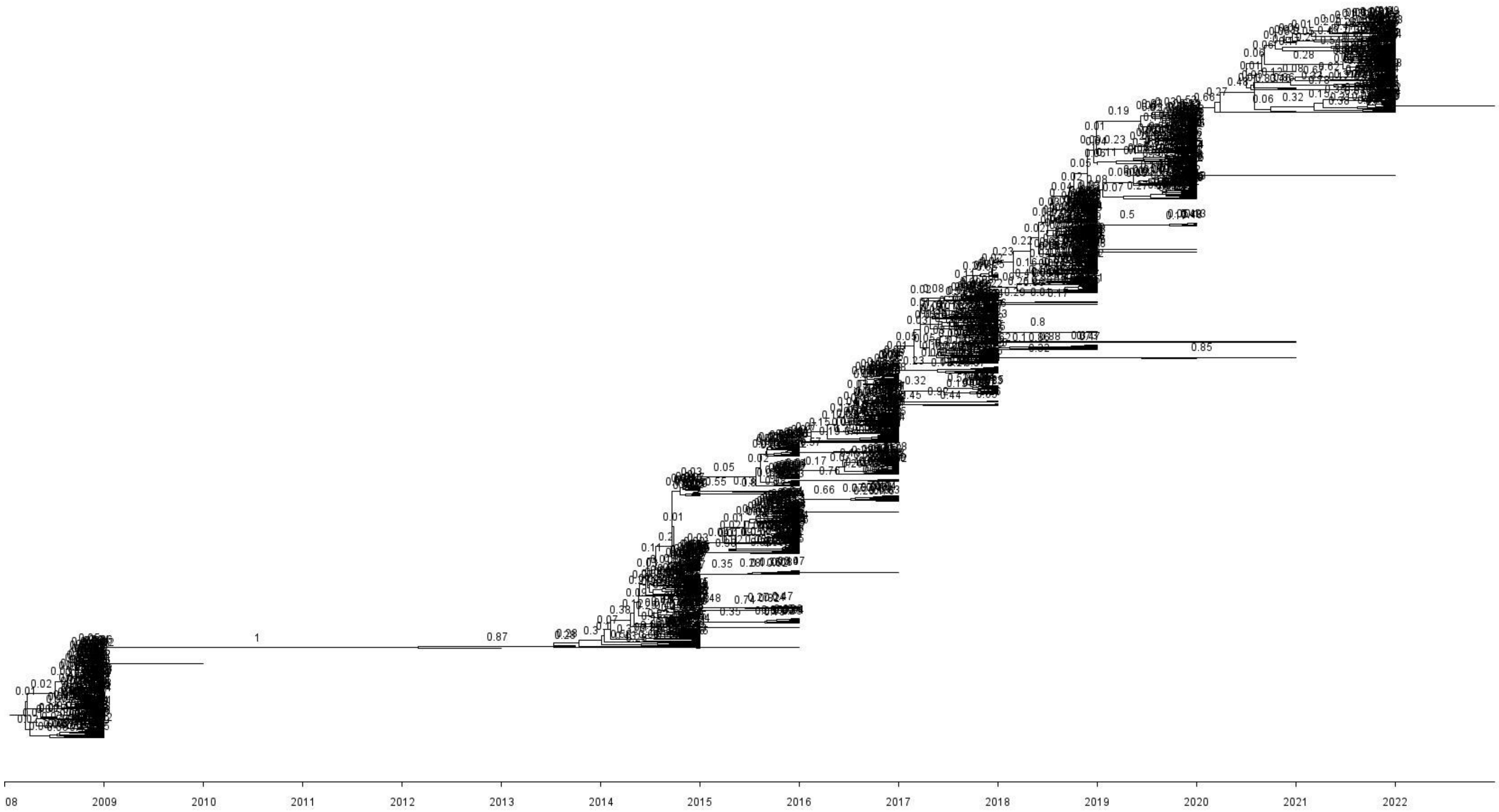
Supplementary Figure 18. Asia NA timetree 2 with posterior probabilities



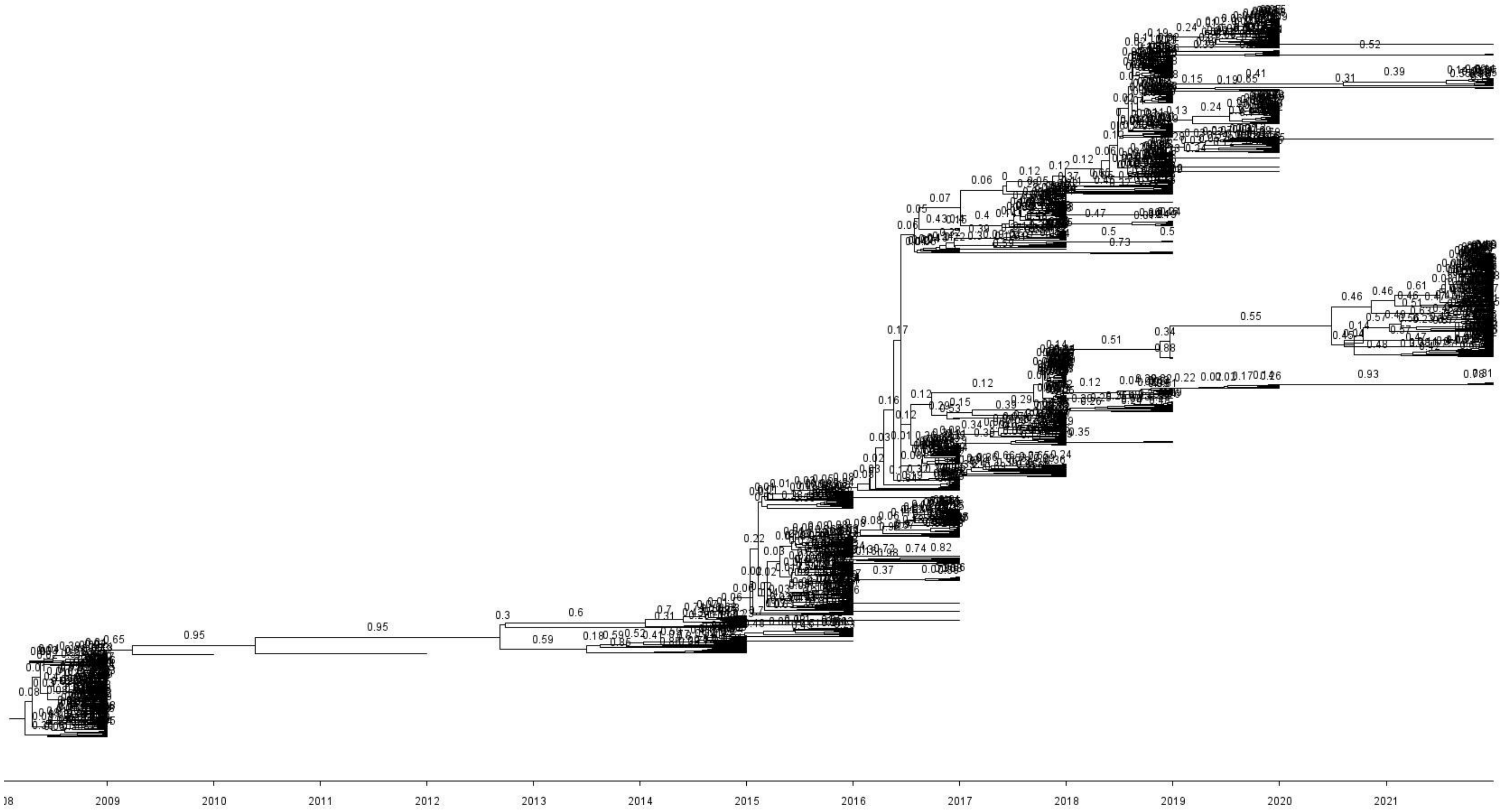
Supplementary Figure 19. Asia NA timetree 3 with posterior probabilities



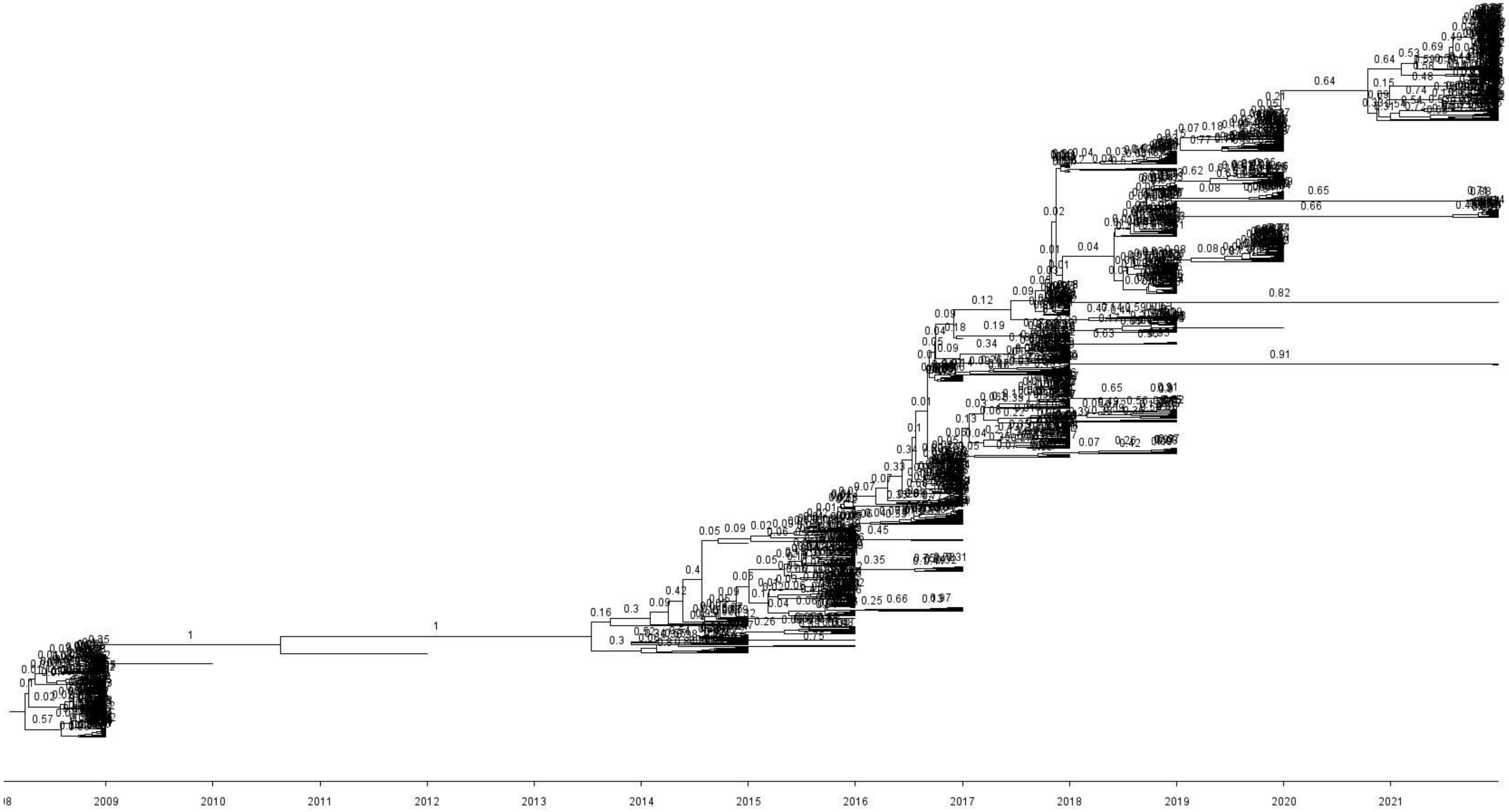
Supplementary Figure 20. Asia NA timetree 4 with posterior probabilities



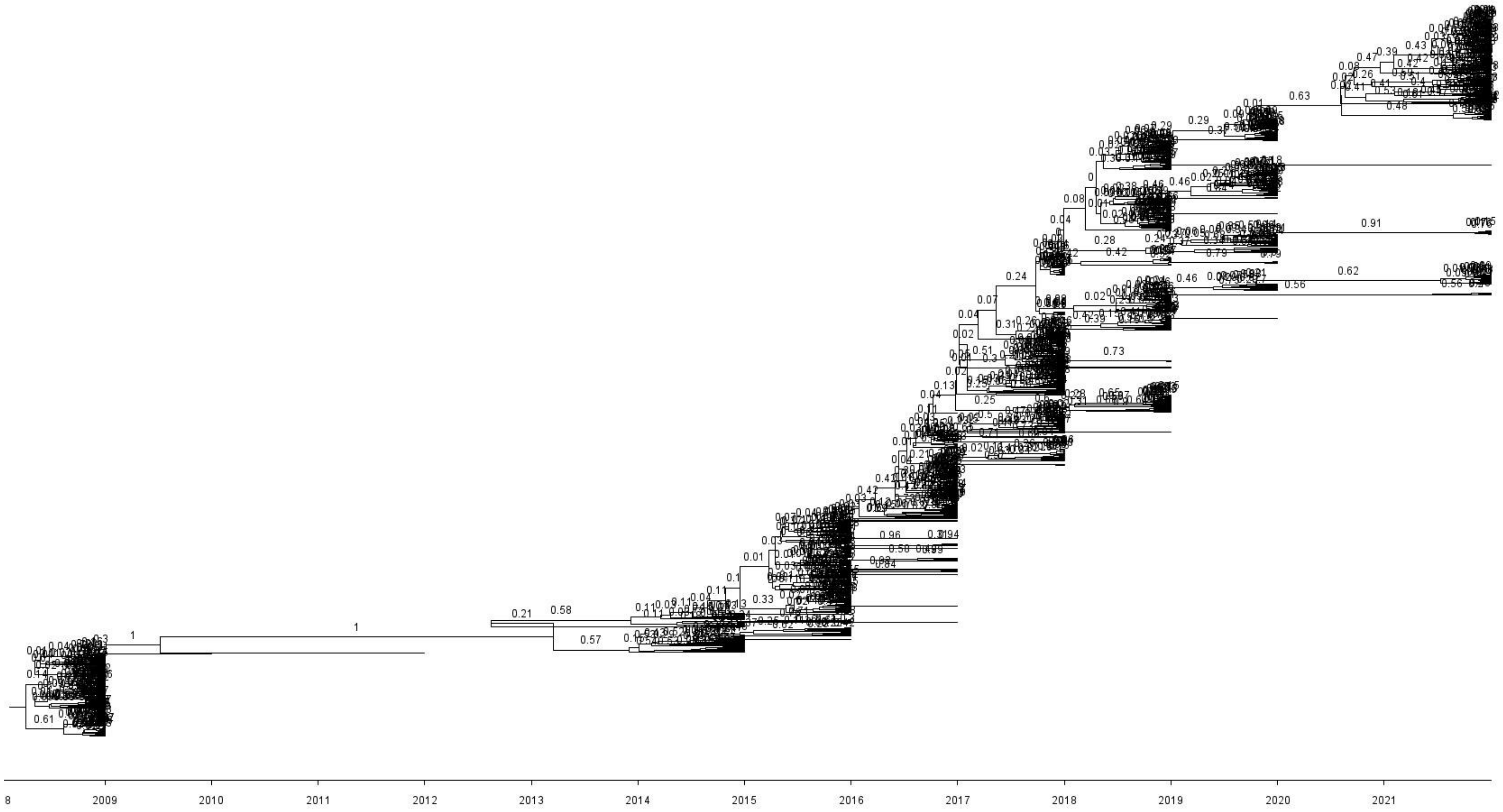
Supplementary Figure 21. Asia NA timetree 5 with posterior probabilities



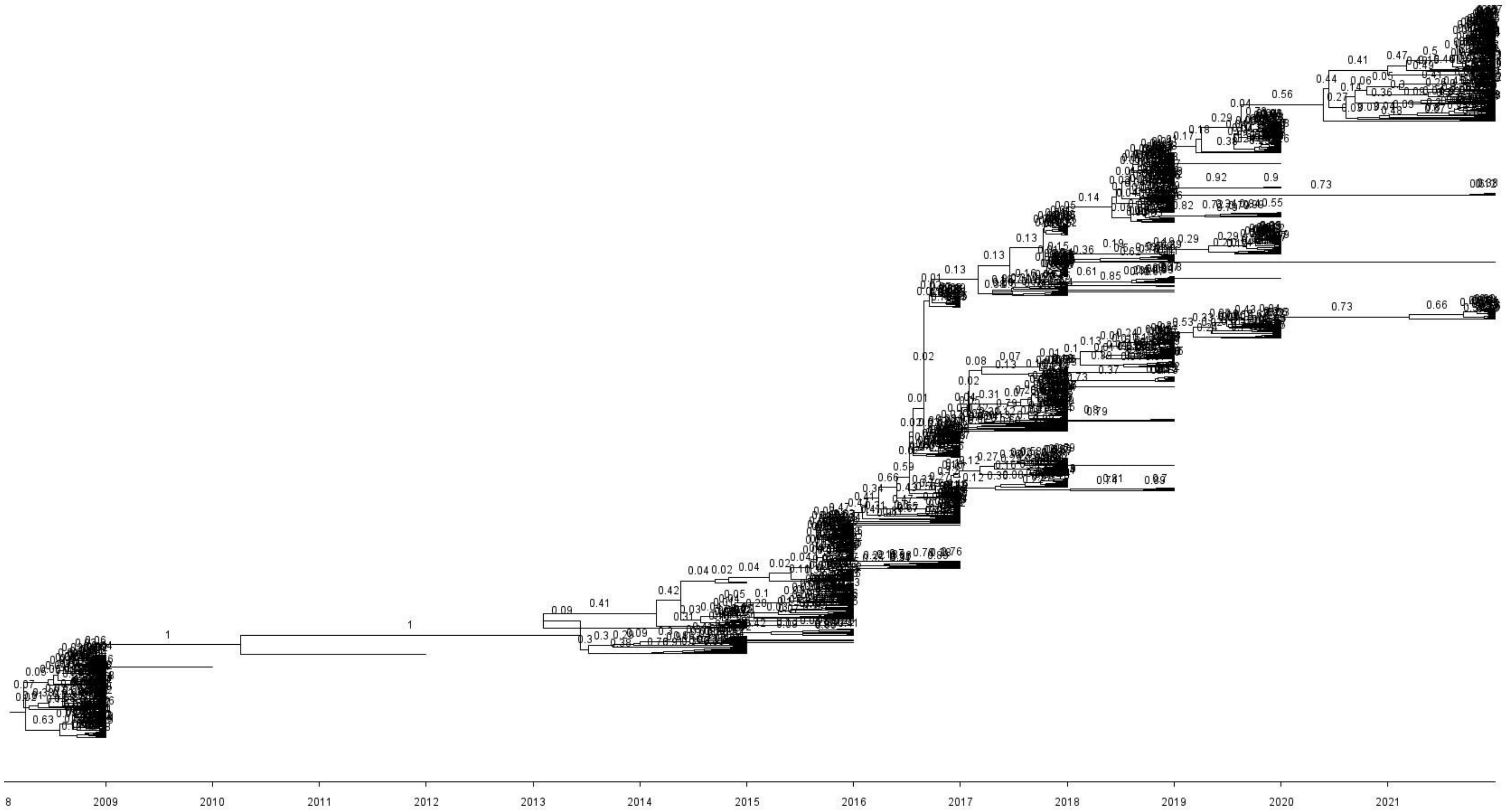
Supplementary Figure 22. Australia HA timetree 1 with posterior probabilities



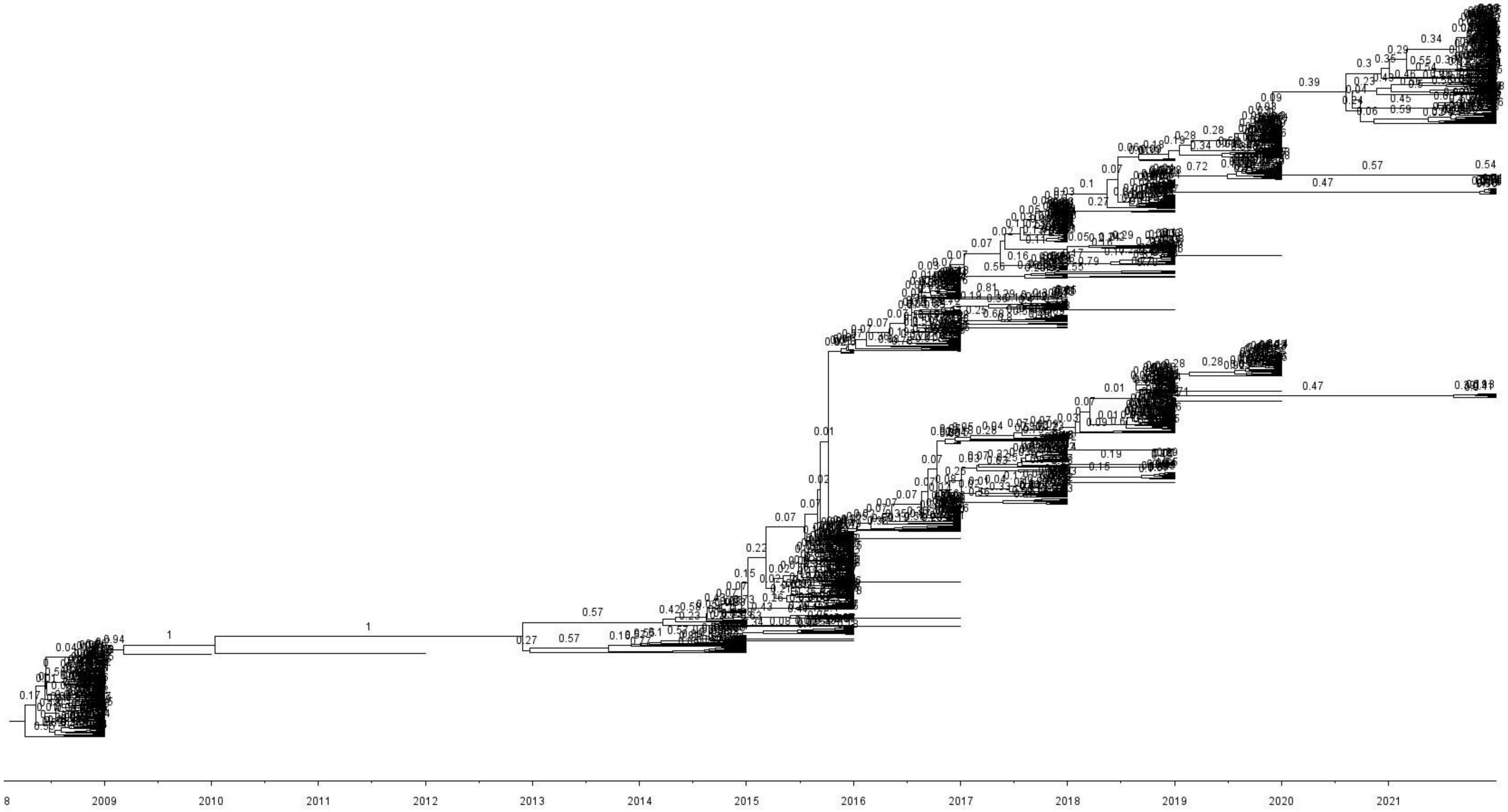
Supplementary Figure 23. Australia HA timetree 2 with posterior probabilities



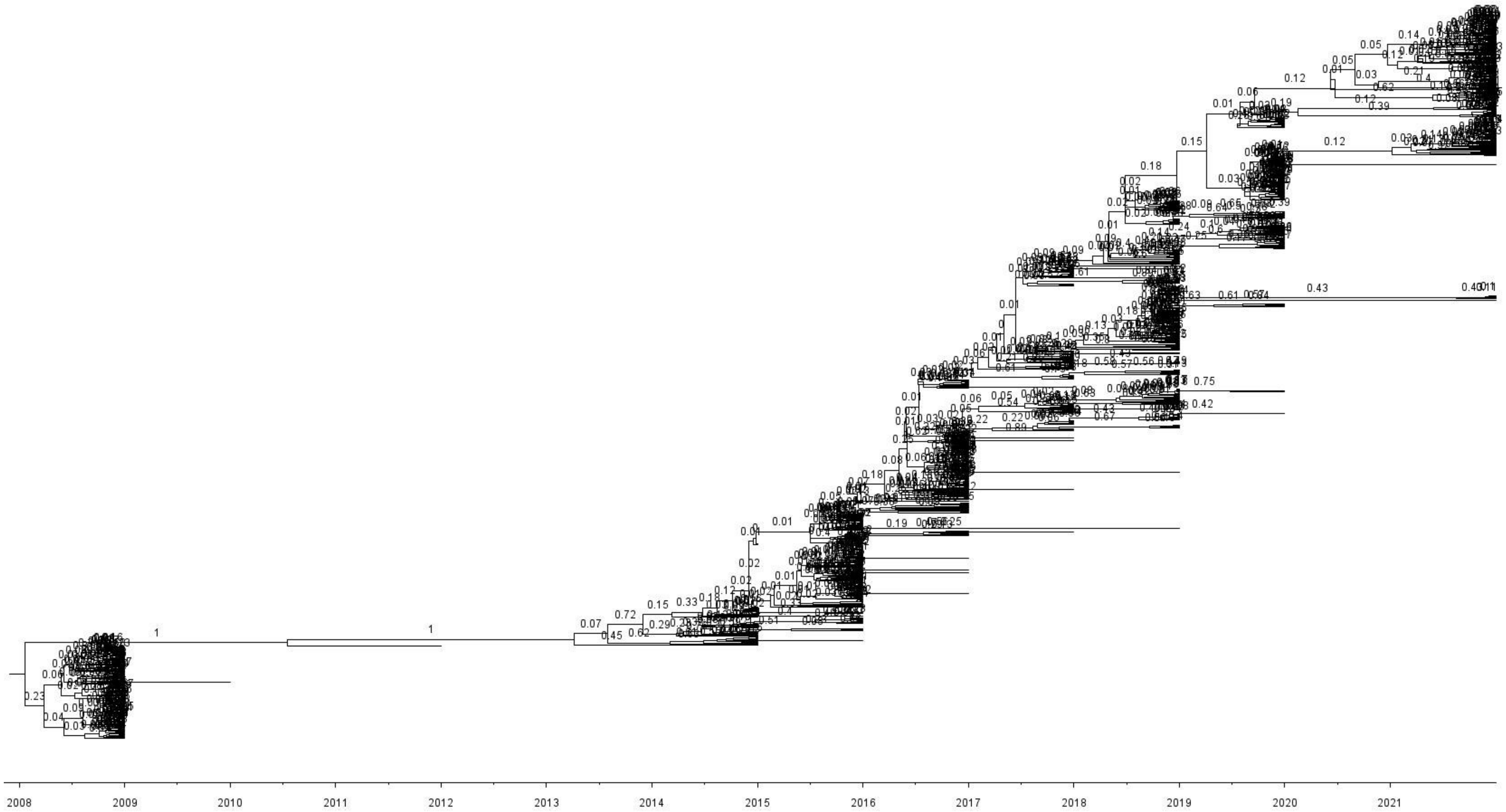
Supplementary Figure 24. Australia HA timetree 3 with posterior probabilities



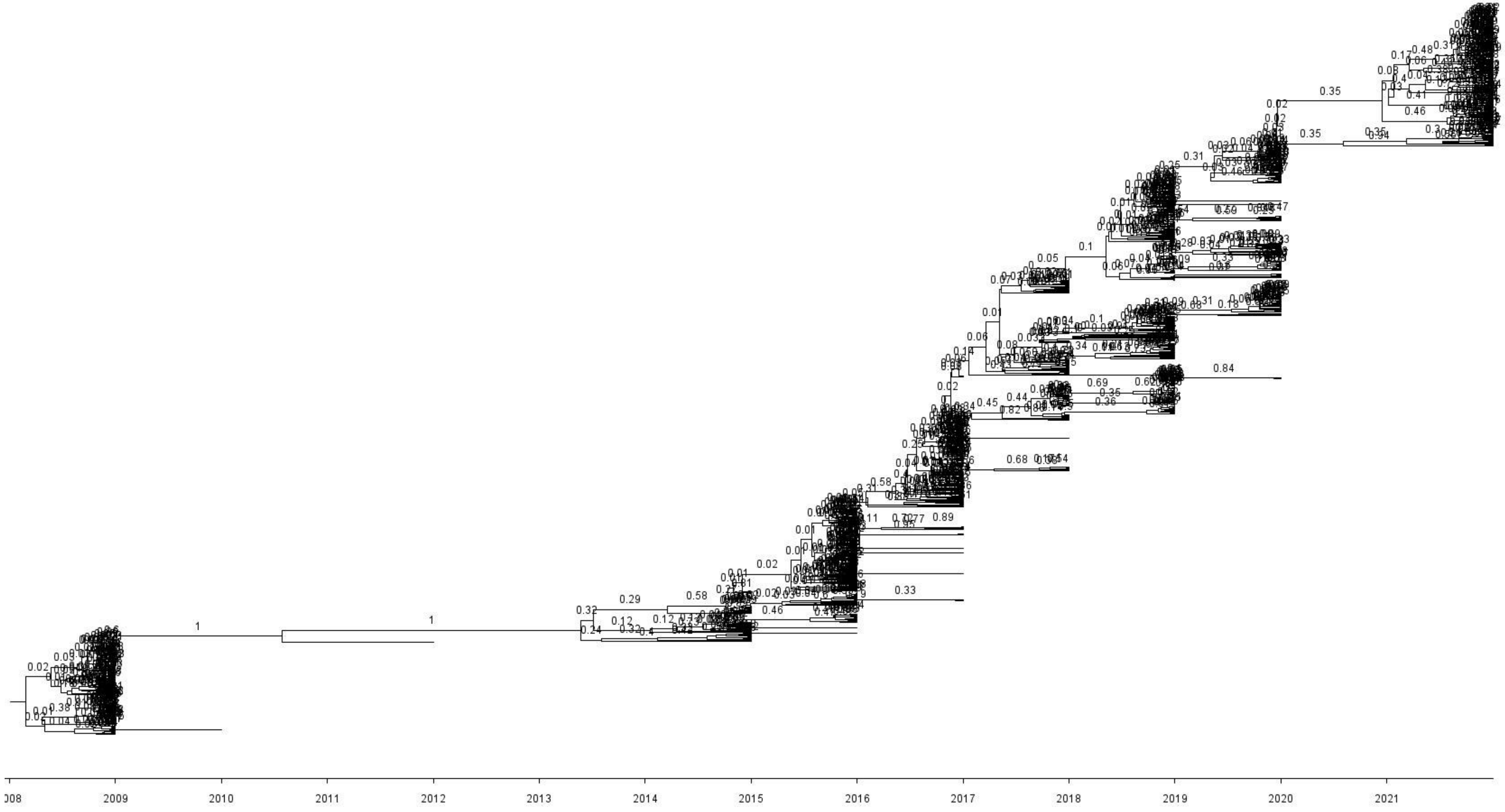
Supplementary Figure 25. Australia HA timetree 4 with posterior probabilities



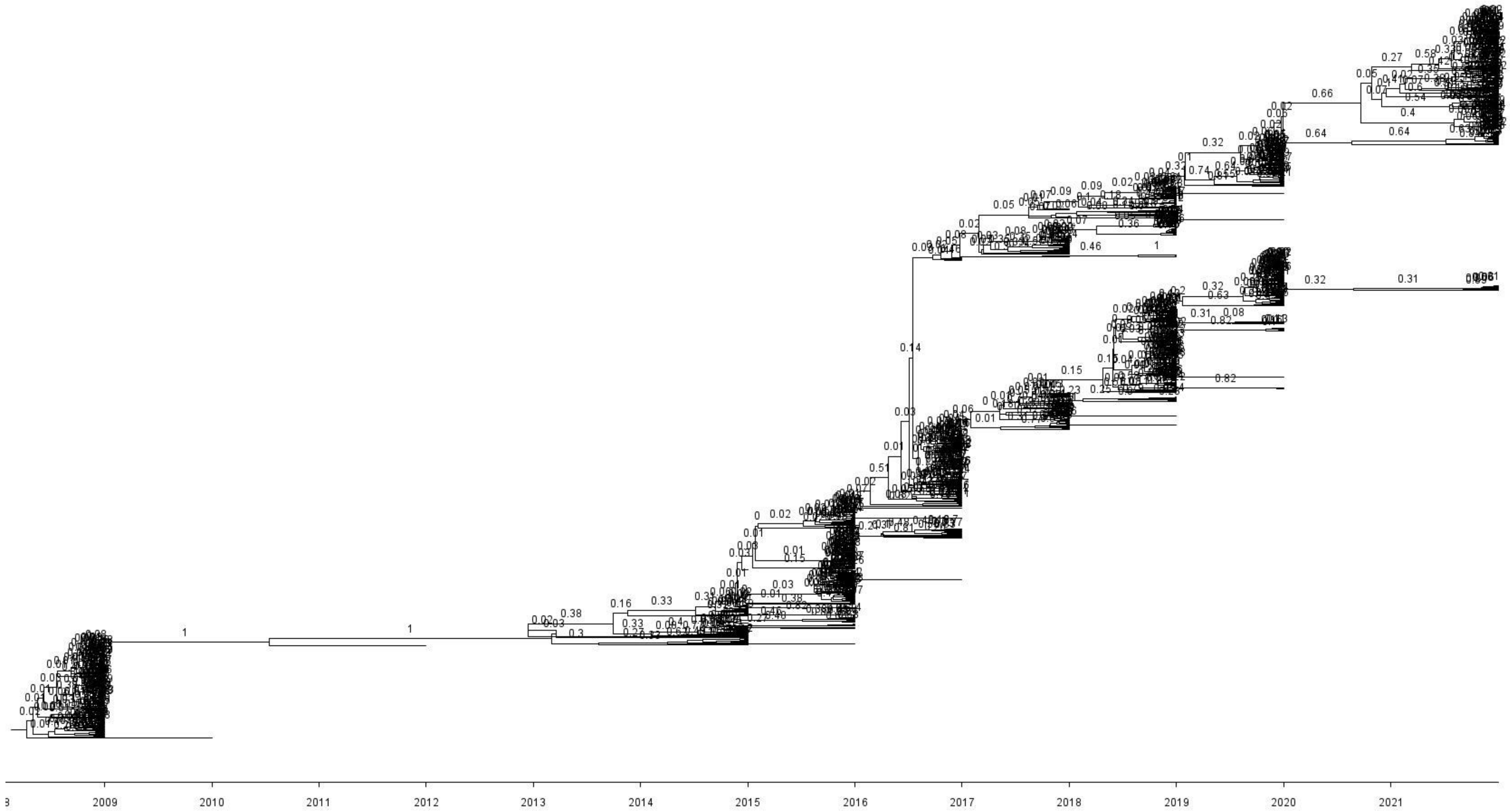
Supplementary Figure 26. Australia HA timetree 5 with posterior probabilities



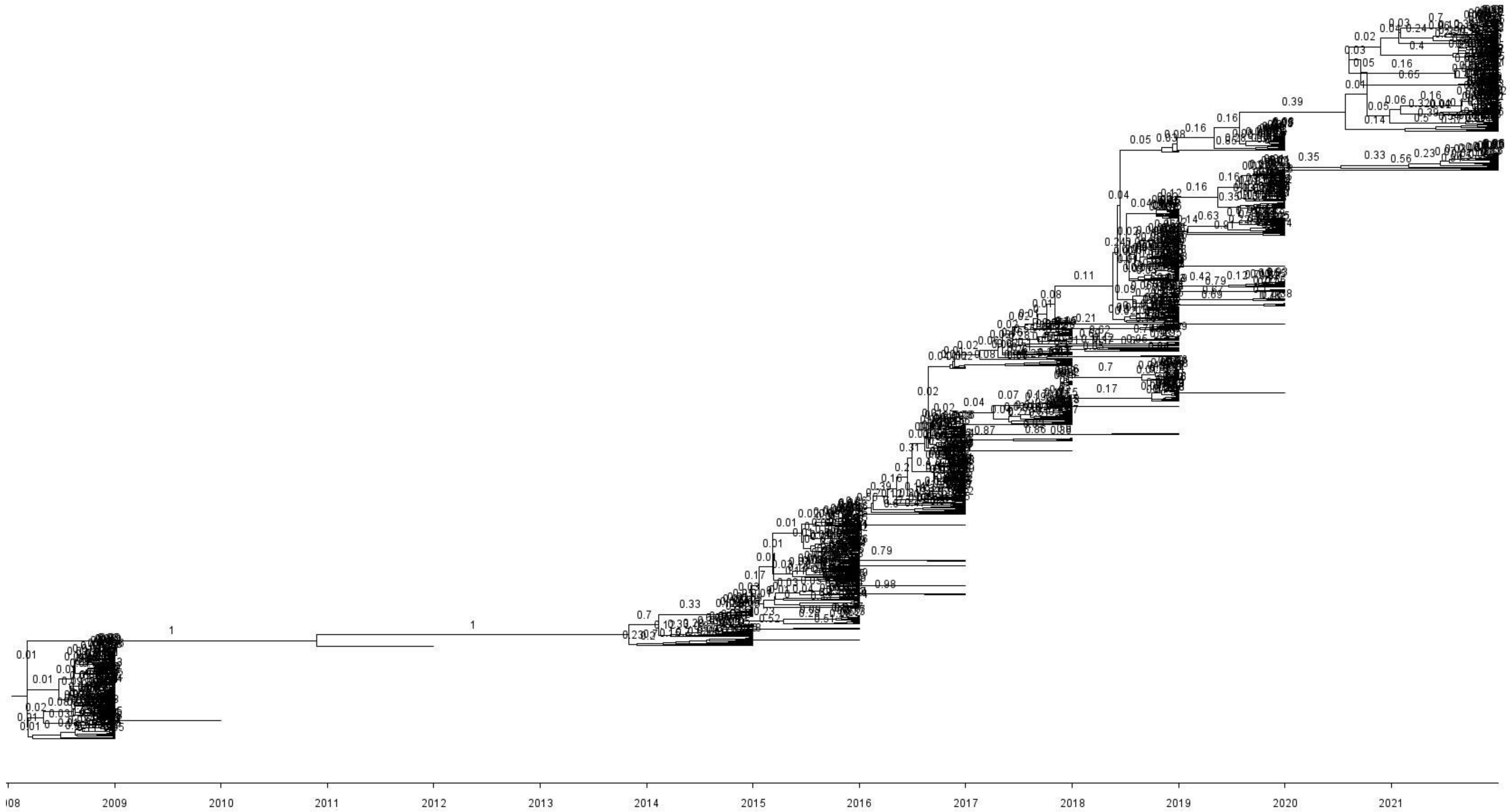
Supplementary Figure 27. Australia NA timetree 1 with posterior probabilities



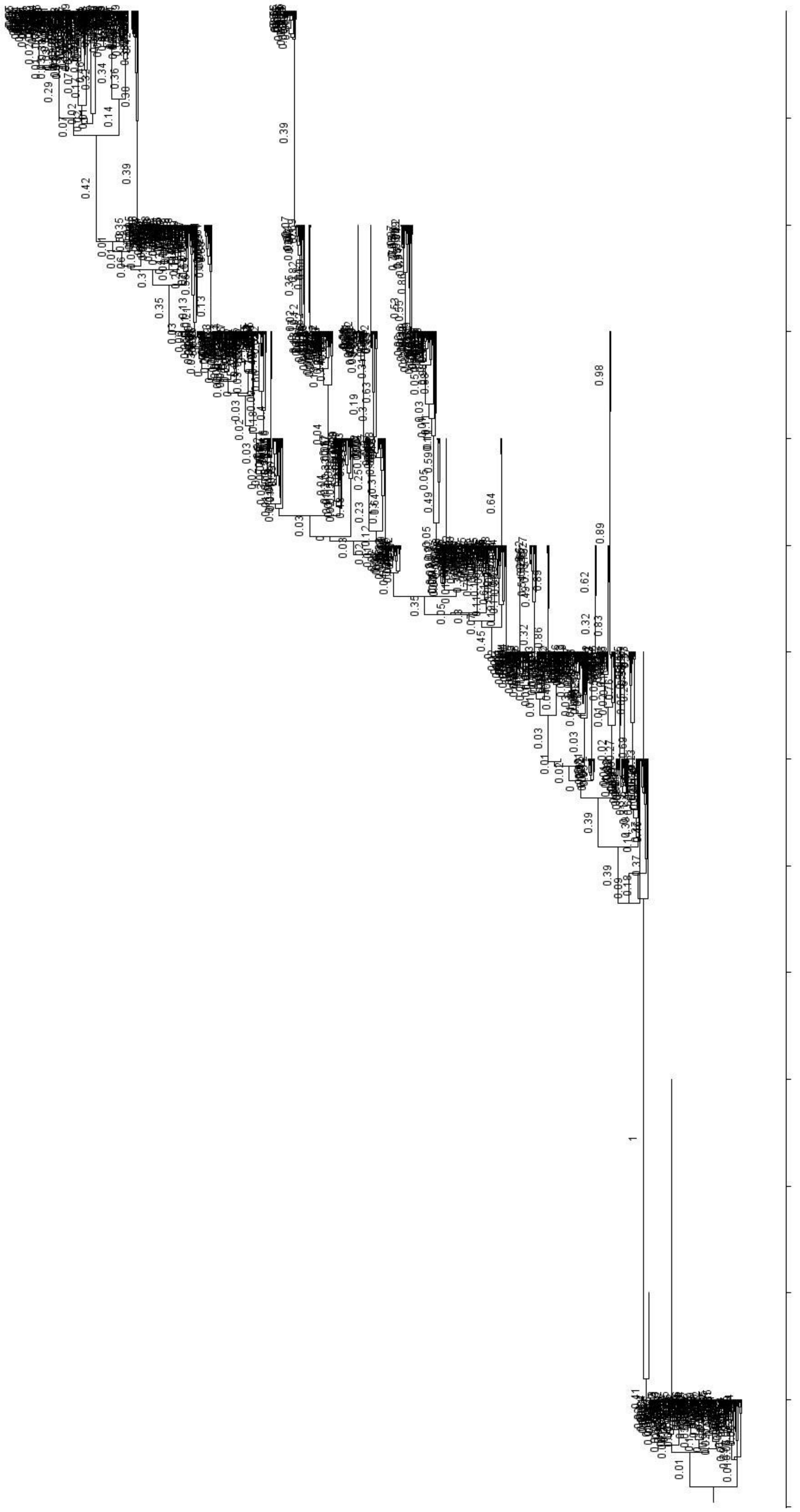
Supplementary Figure 28. Australia NA timetree 2 with posterior probabilities



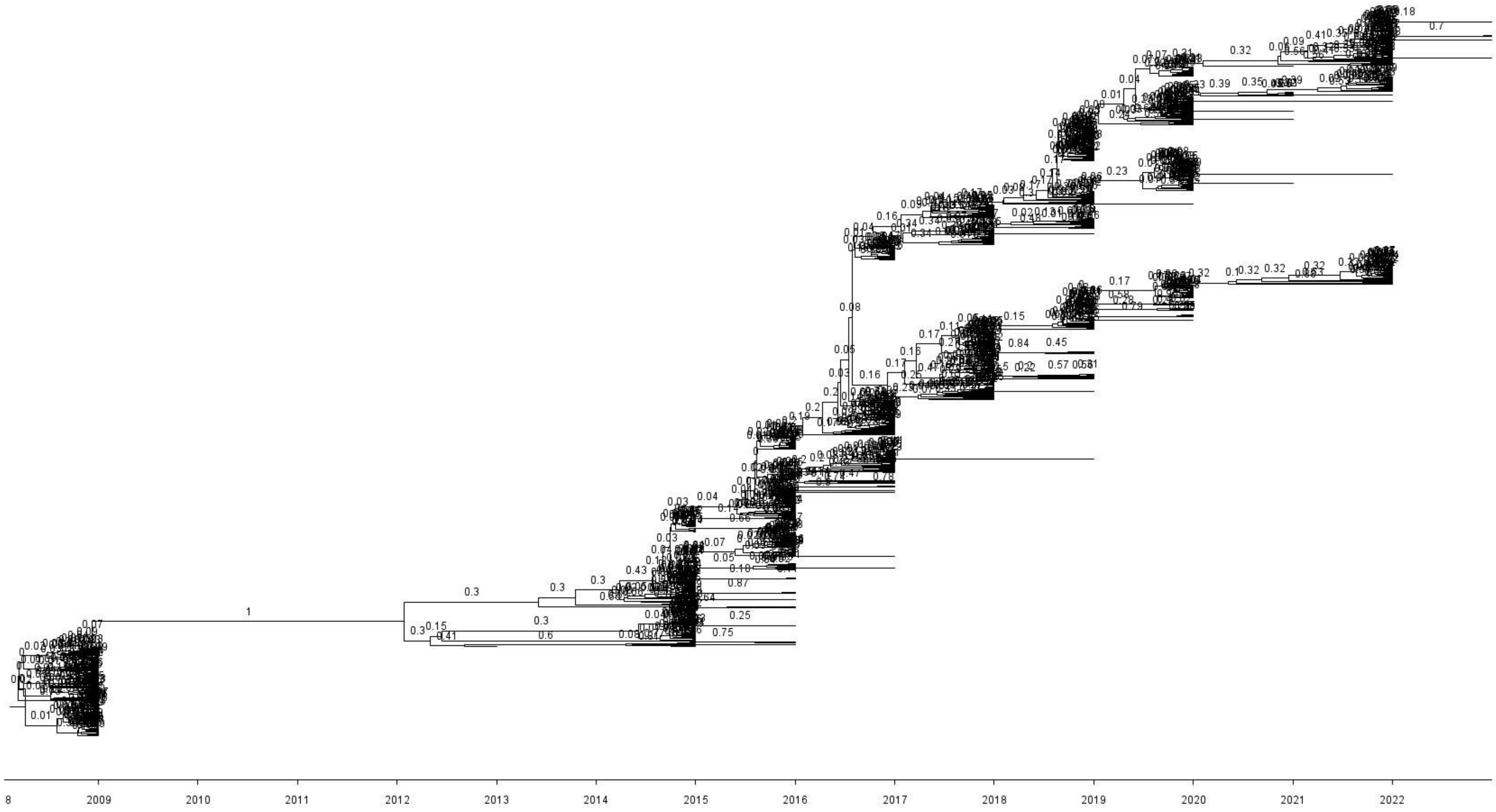
Supplementary Figure 29. Australia NA timetree 3 with posterior probabilities



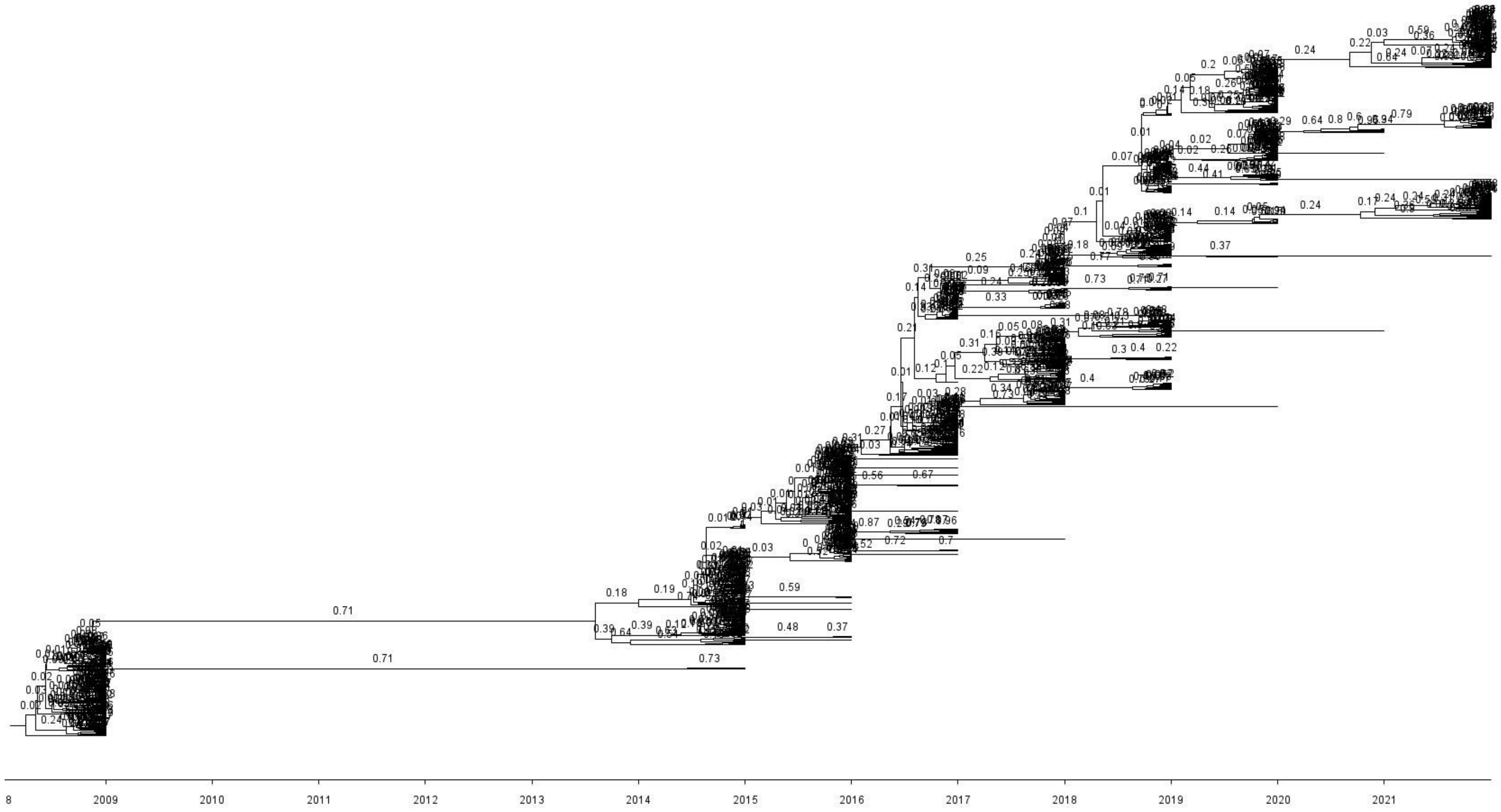
Supplementary Figure 30. Australia NA timetree 4 with posterior probabilities



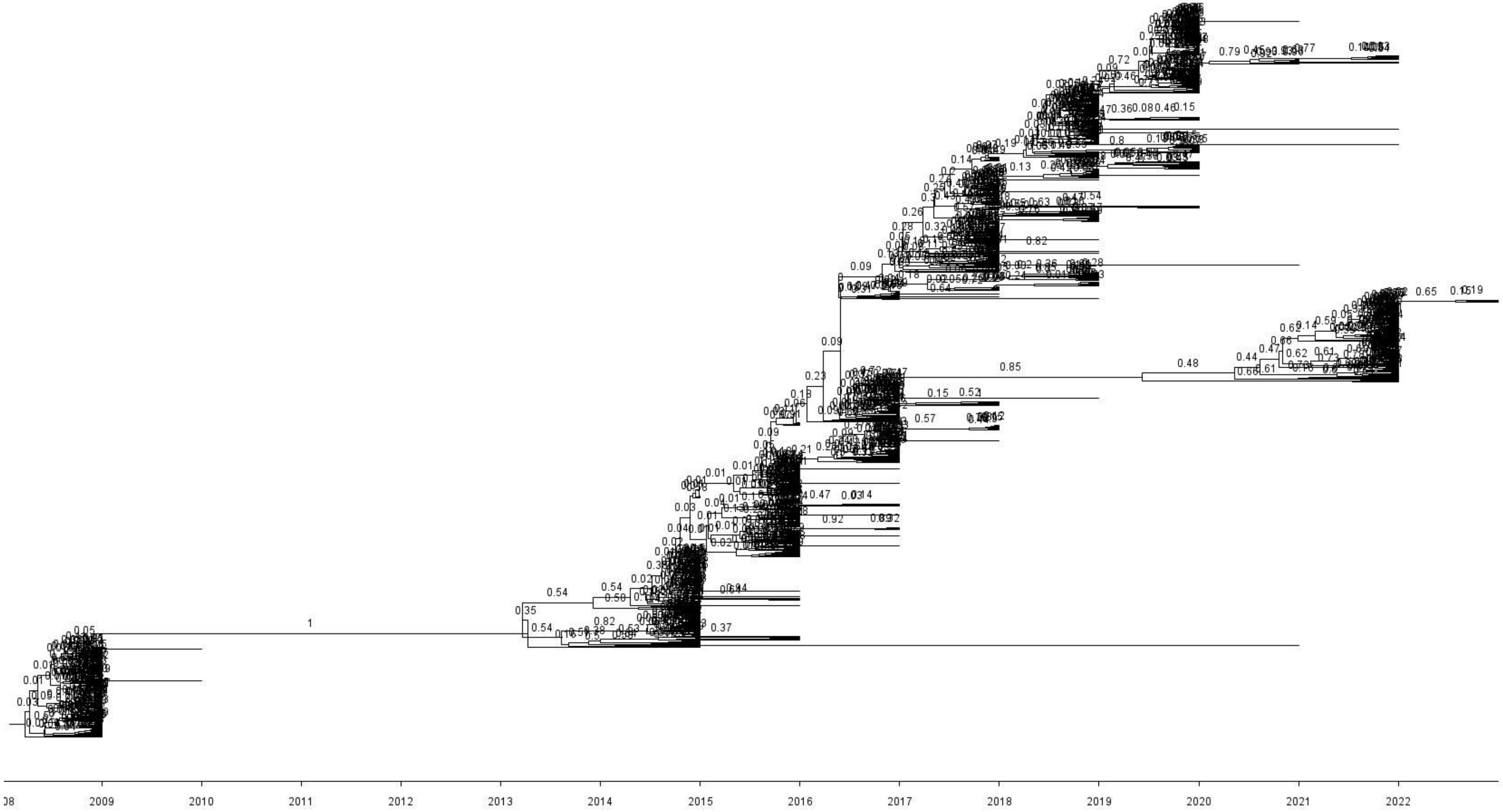
Supplementary Figure 31. Australia NA timetree 5 with posterior probabilities



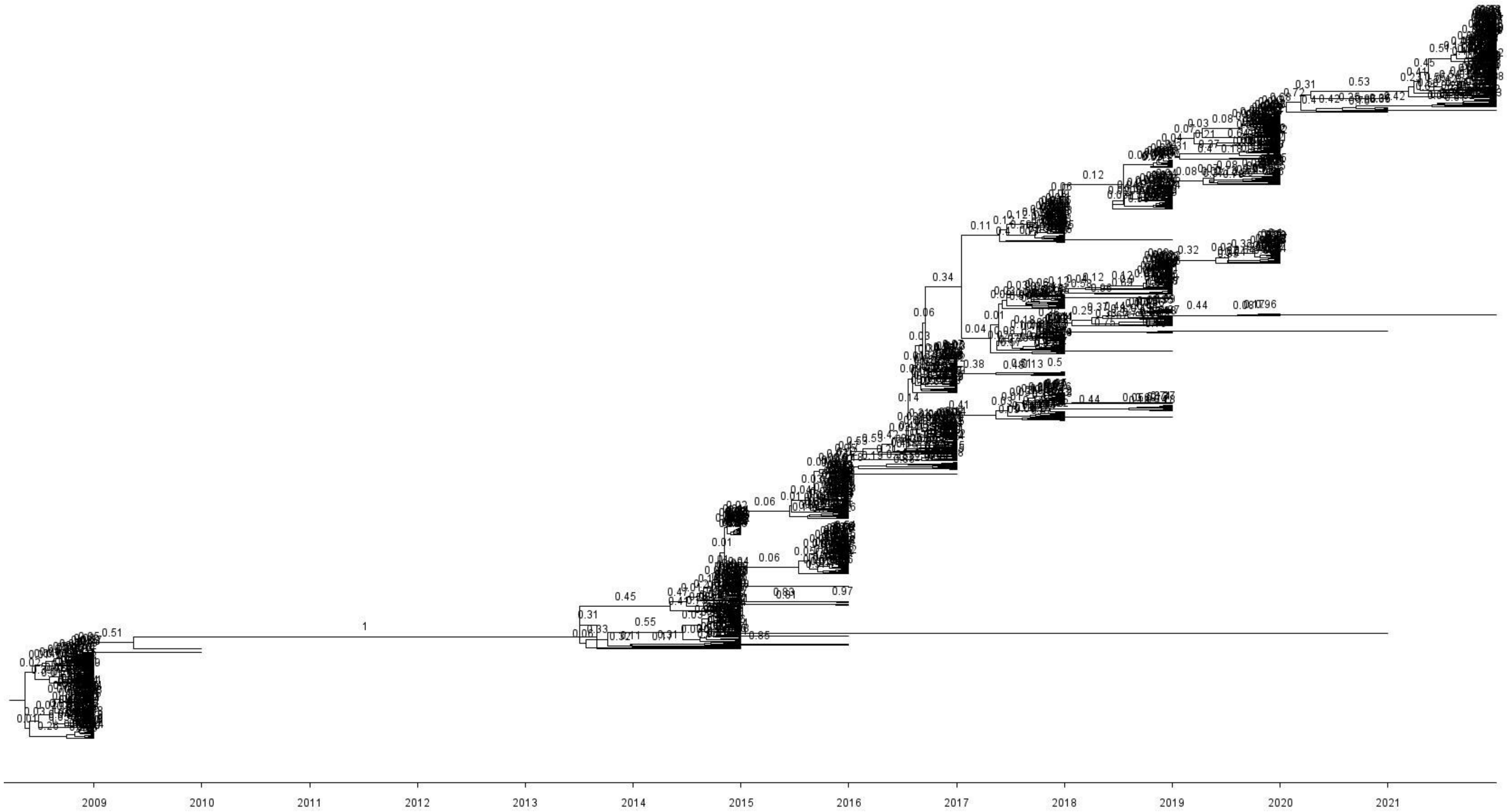
Supplementary Figure 32. Europe HA timetree 1 with posterior probabilities



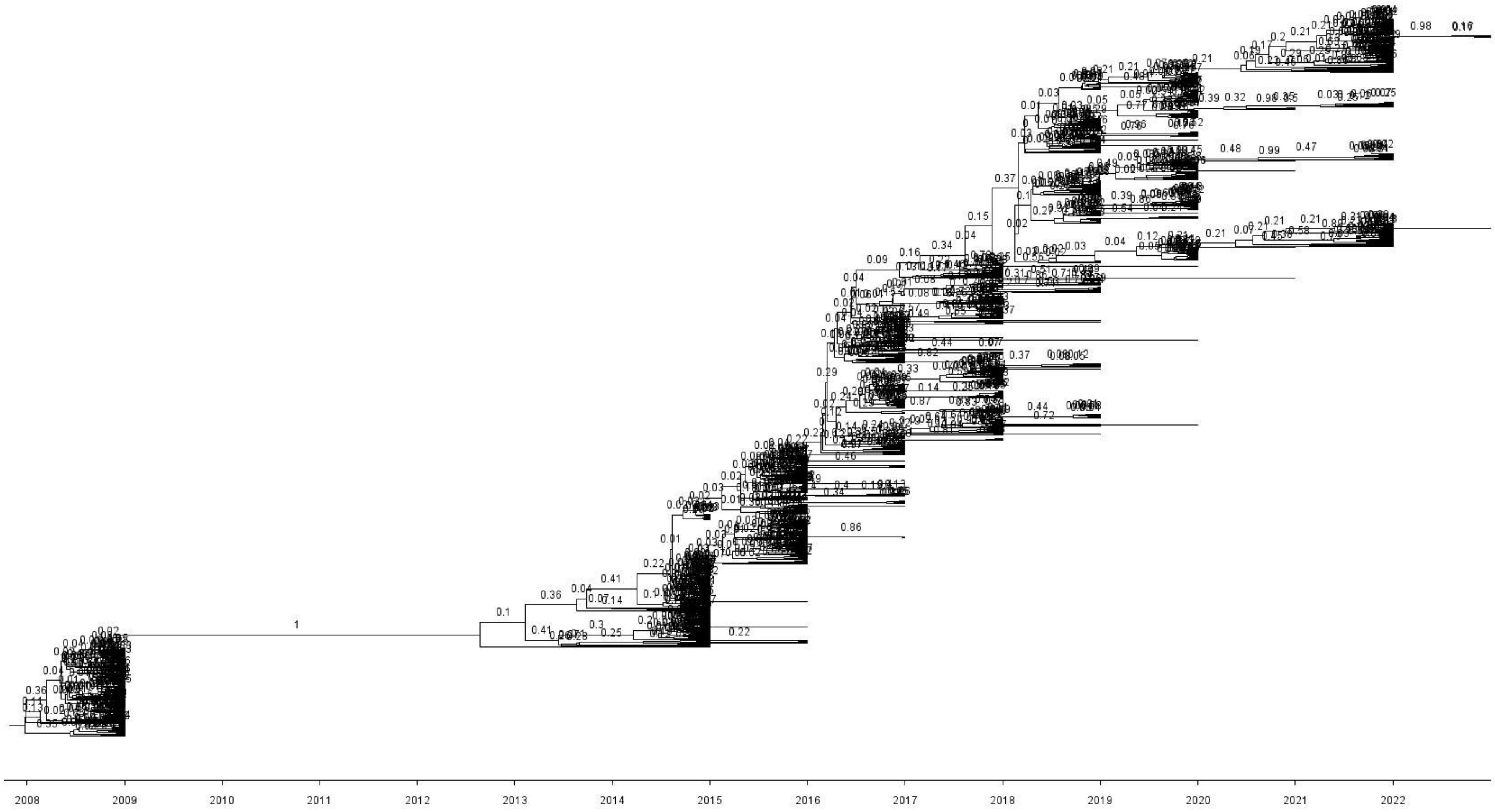
Supplementary Figure 33. Europe HA timetree 2 with posterior probabilities



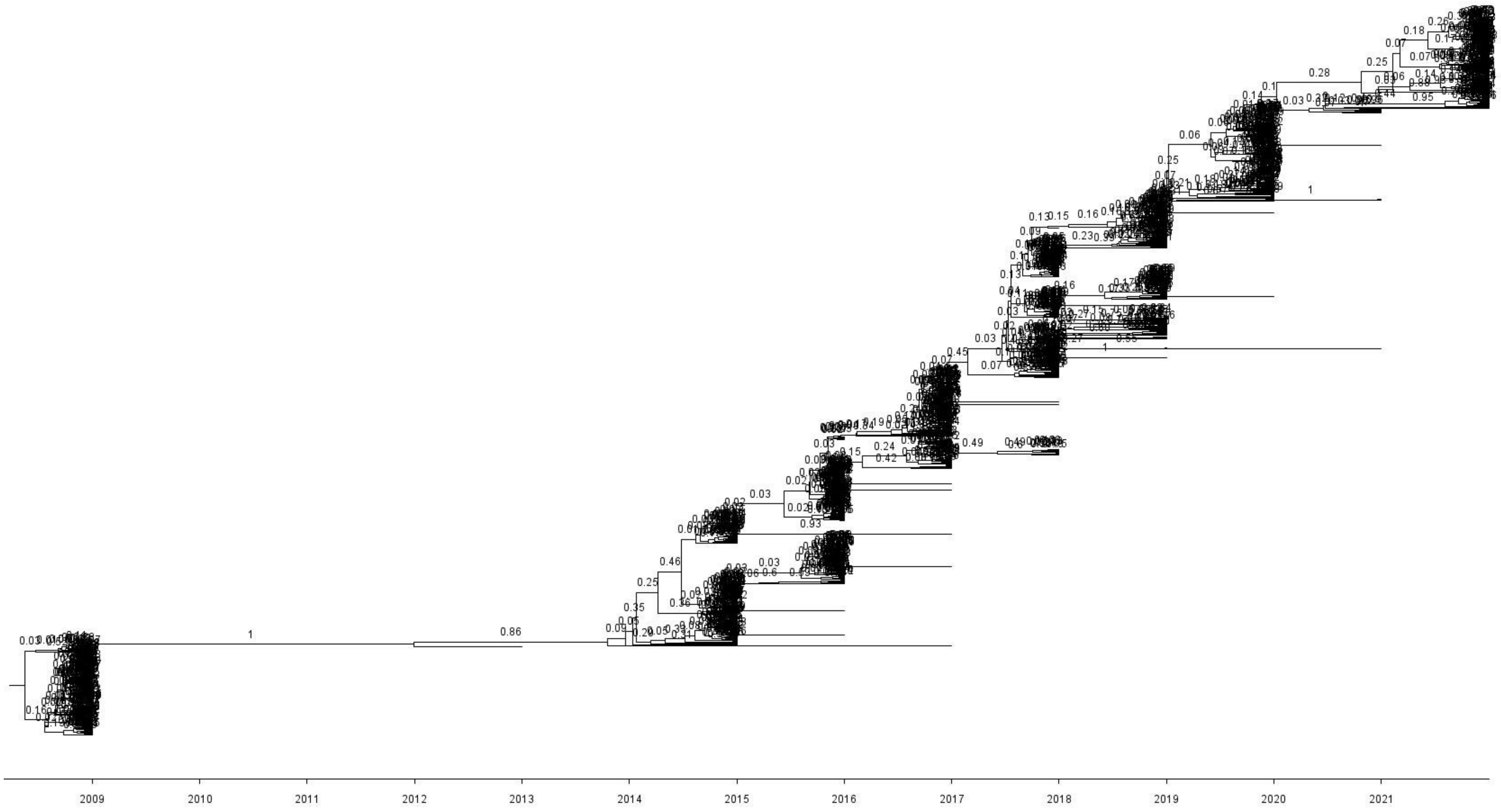
Supplementary Figure 34. Europe HA timetree 3 with posterior probabilities



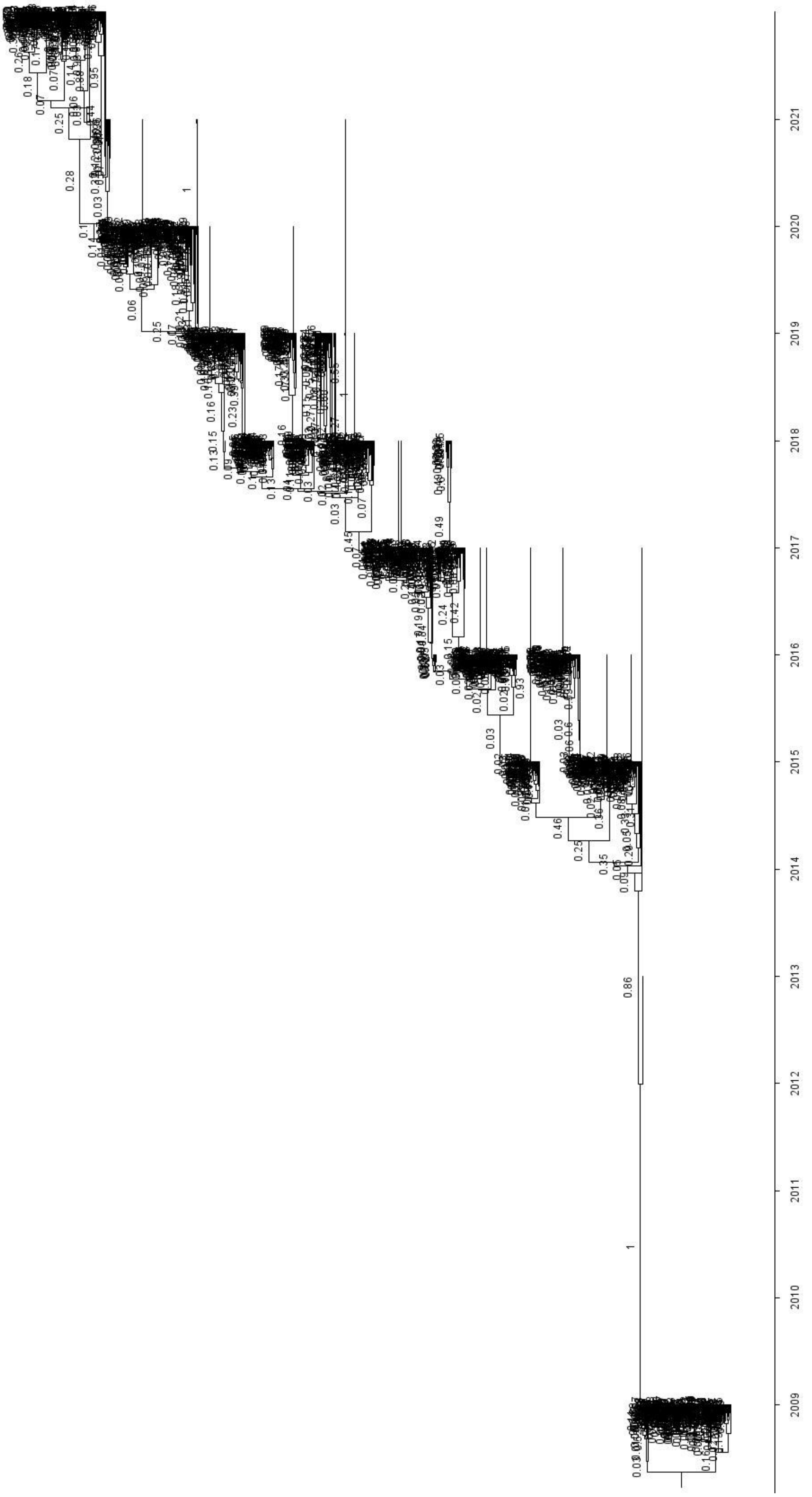
Supplementary Figure 35. Europe HA timetree 4 with posterior probabilities



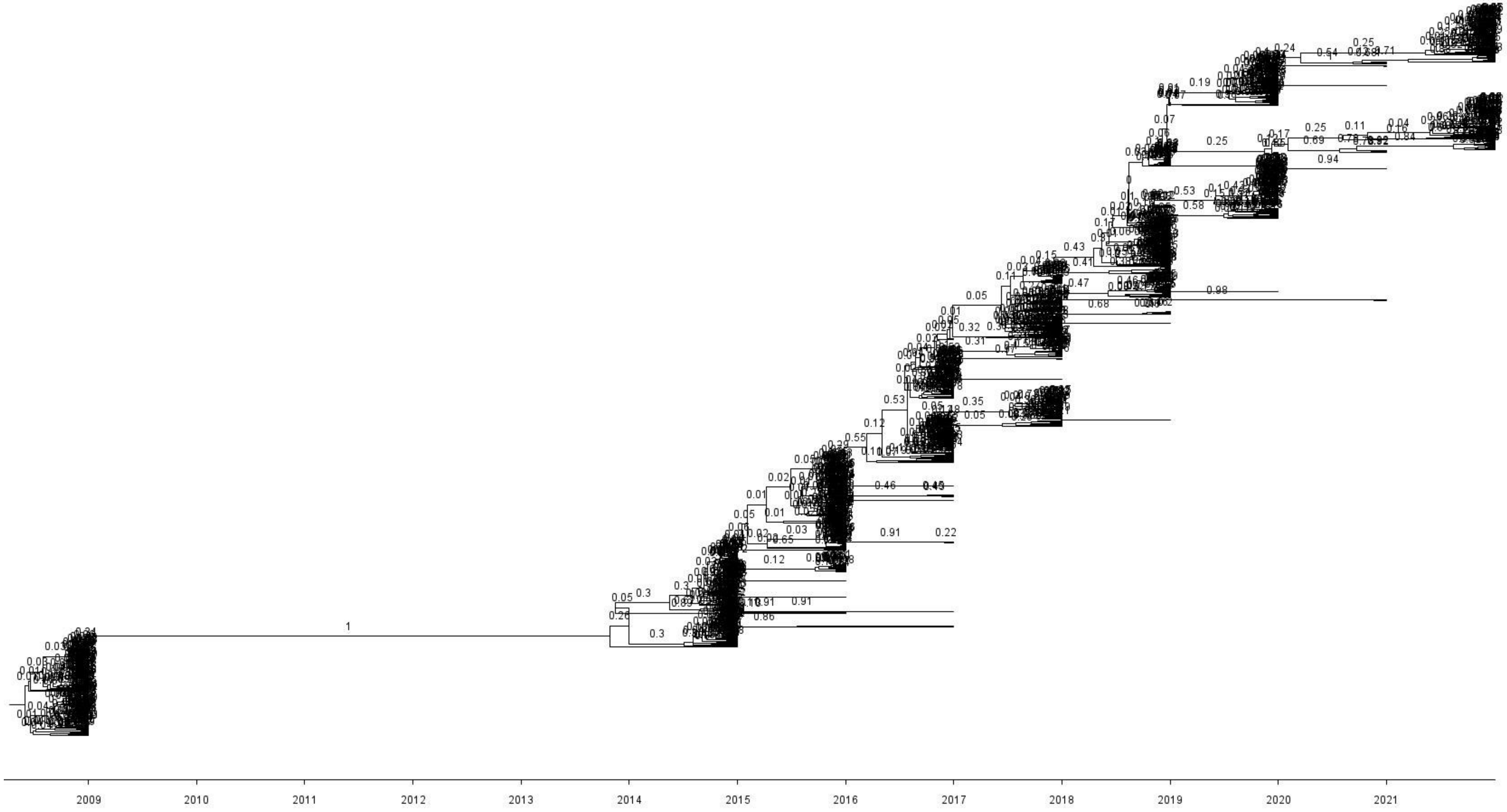
Supplementary Figure 36. Europe HA timetree 5 with posterior probabilities



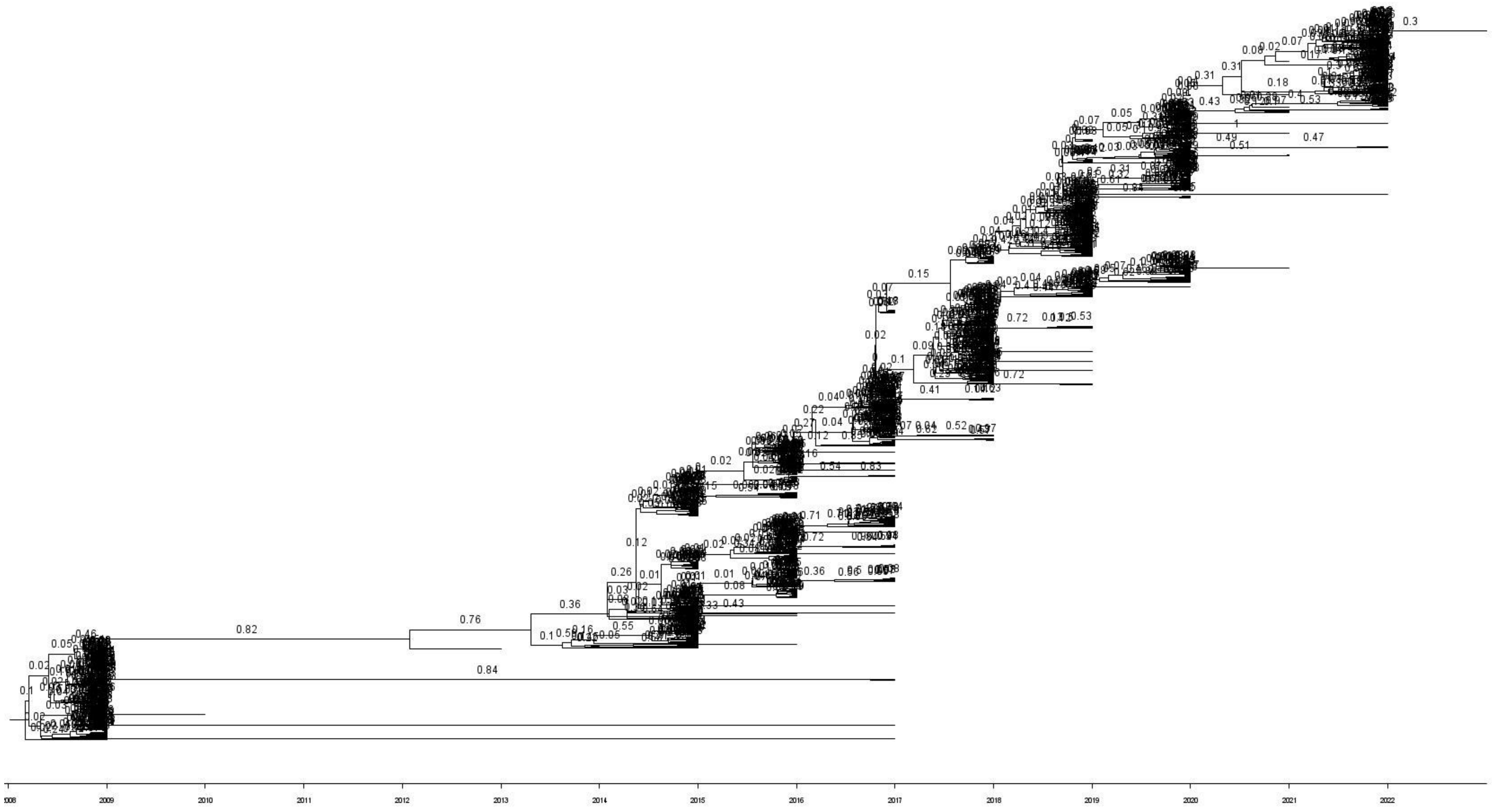
Supplementary Figure 36. Europe HA timetree 5 with posterior probabilities



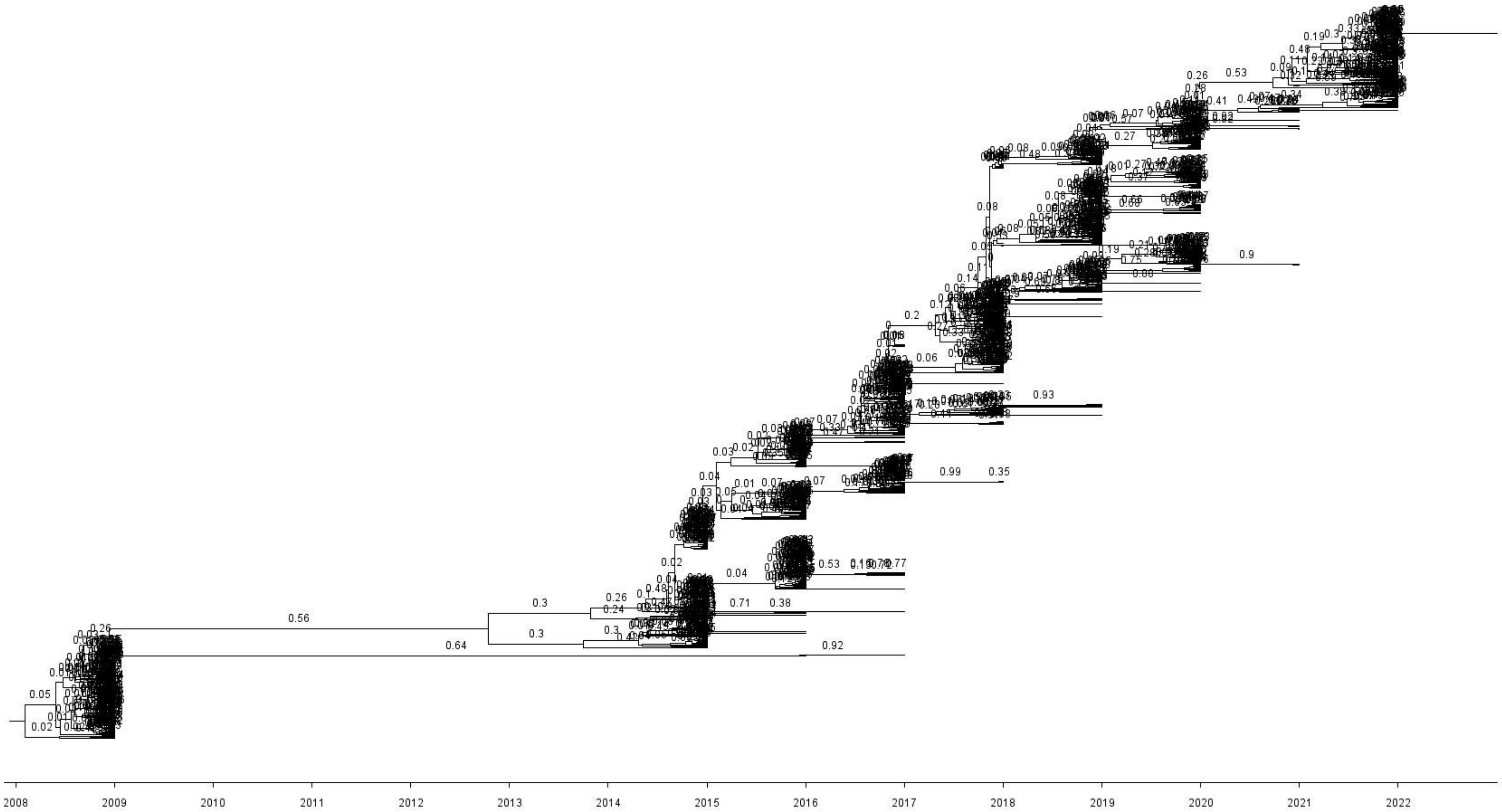
Supplementary Figure 37. Europe NA timetree 1 with posterior probabilities



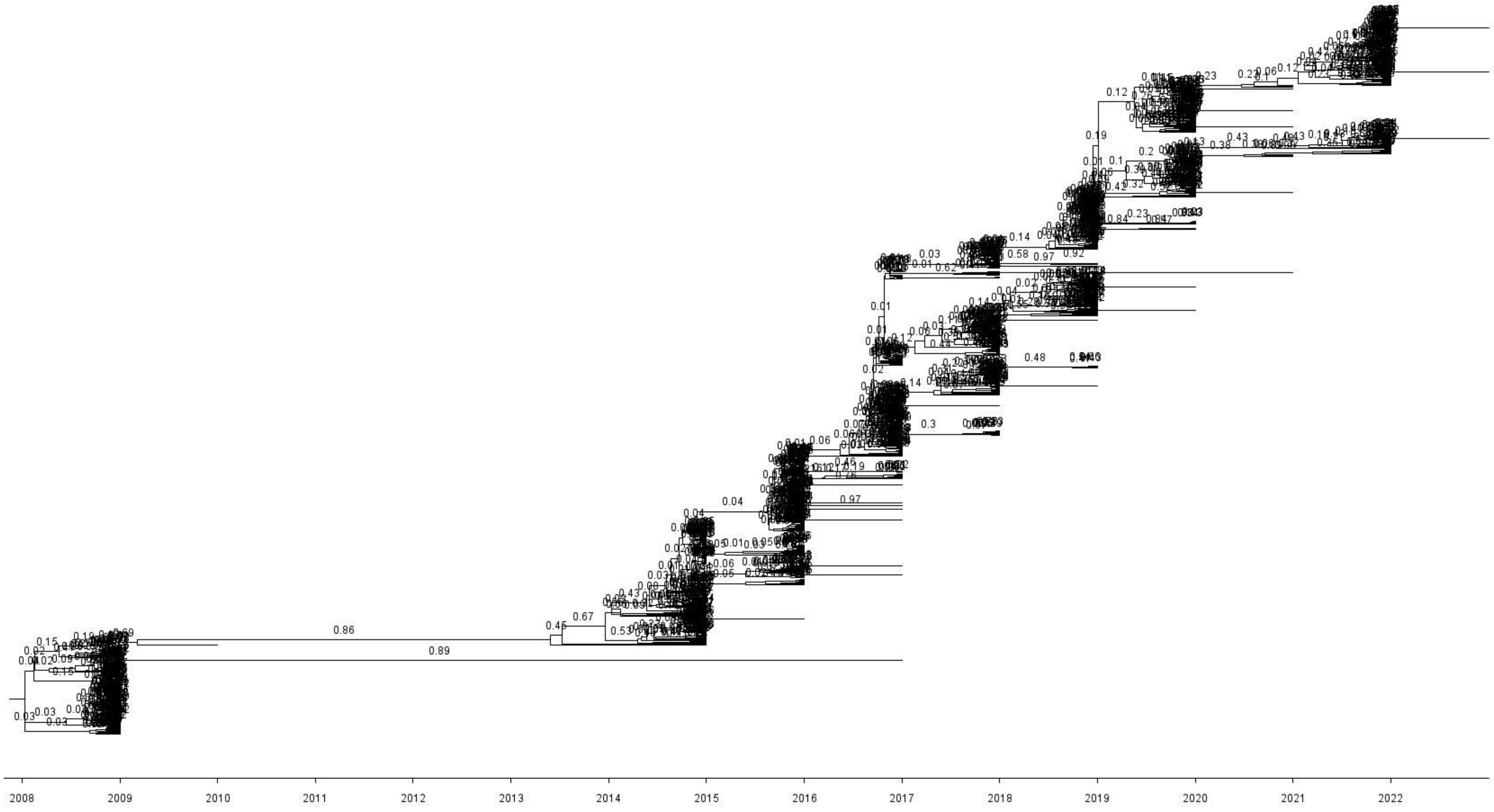
Supplementary Figure 38. Europe NA timetree 2 with posterior probabilities



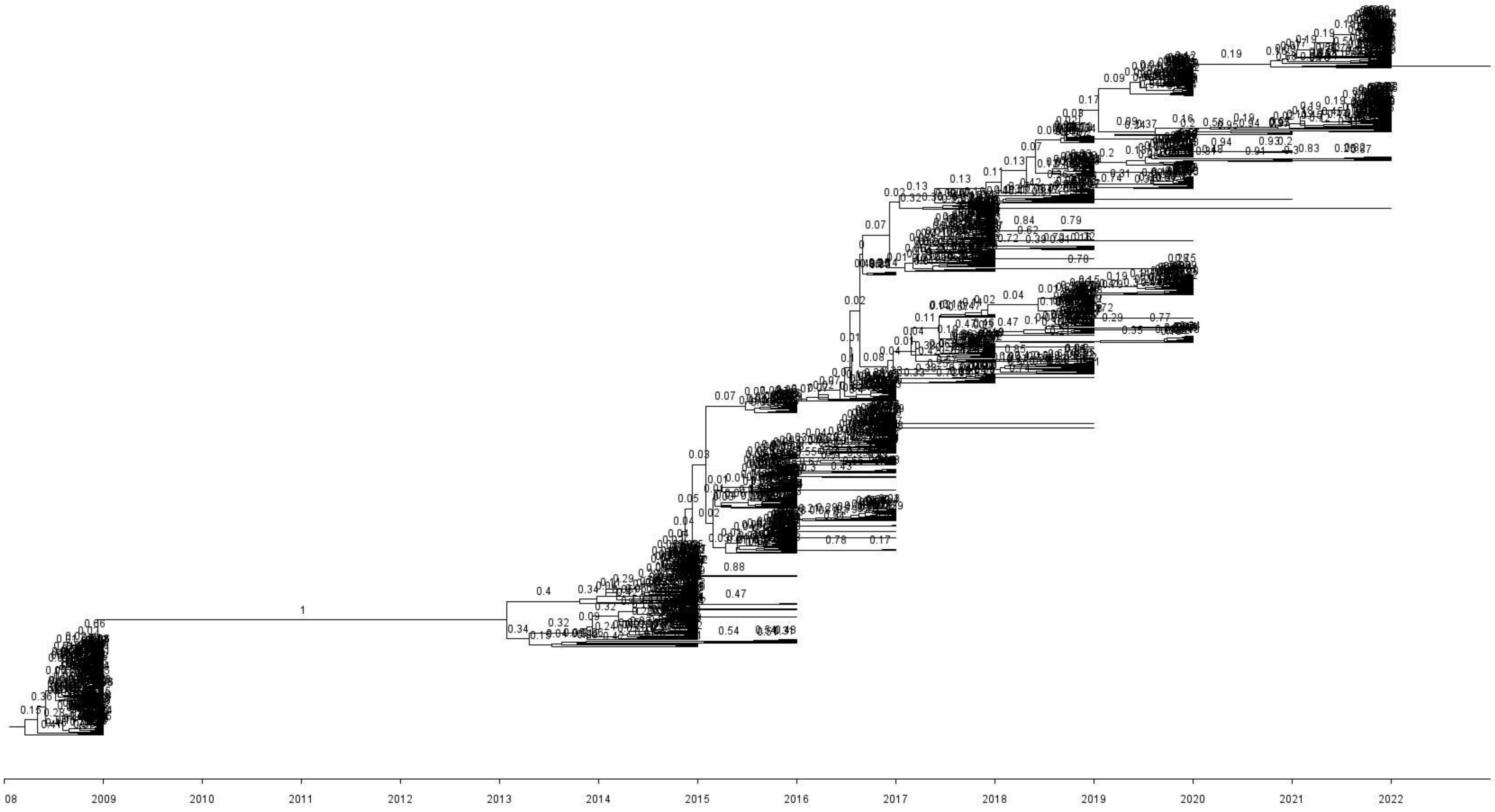
Supplementary Figure 38. Europe NA timetree 3 with posterior probabilities



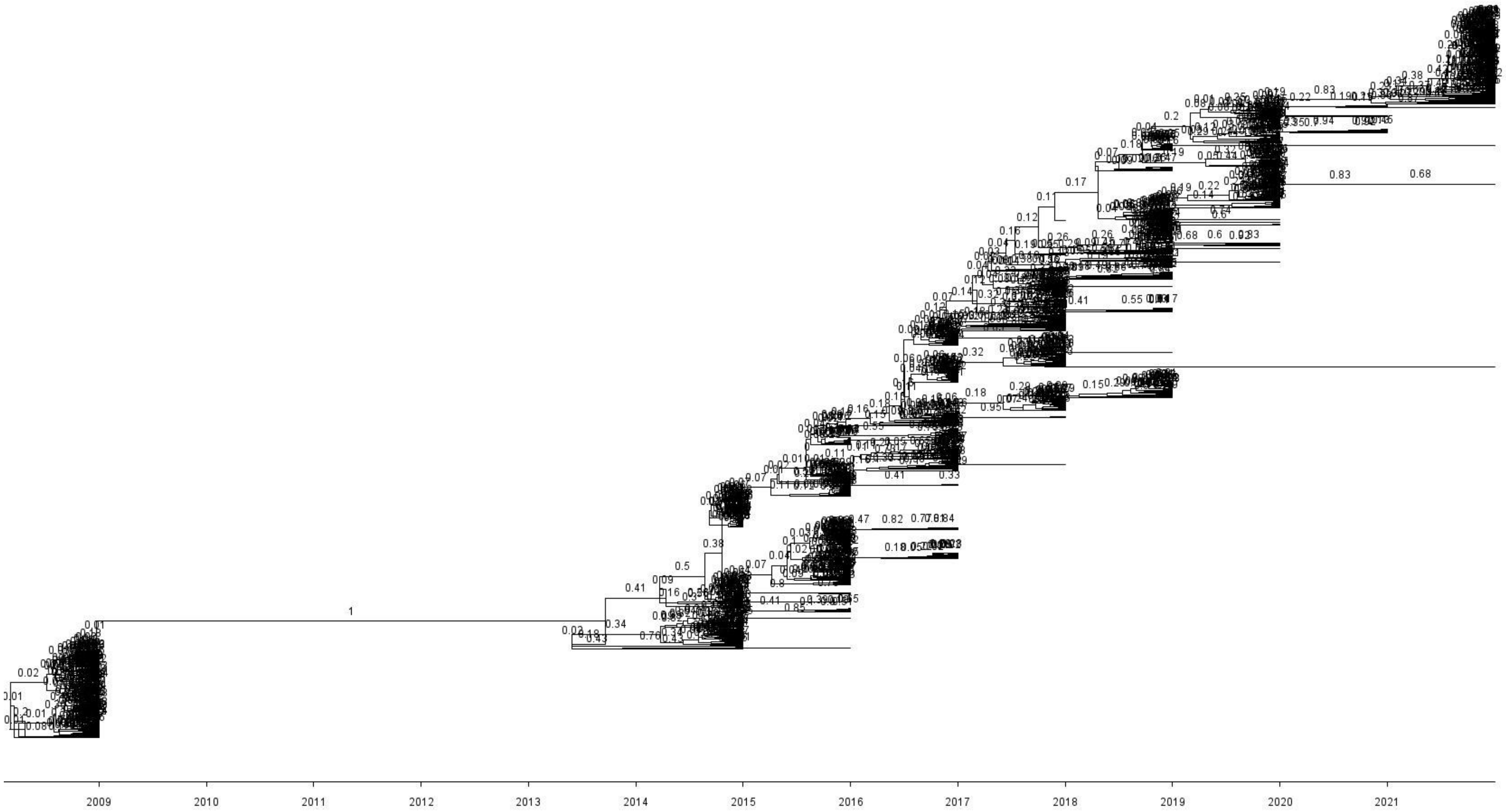
Supplementary Figure 40. Europe NA timetree 4 with posterior probabilities



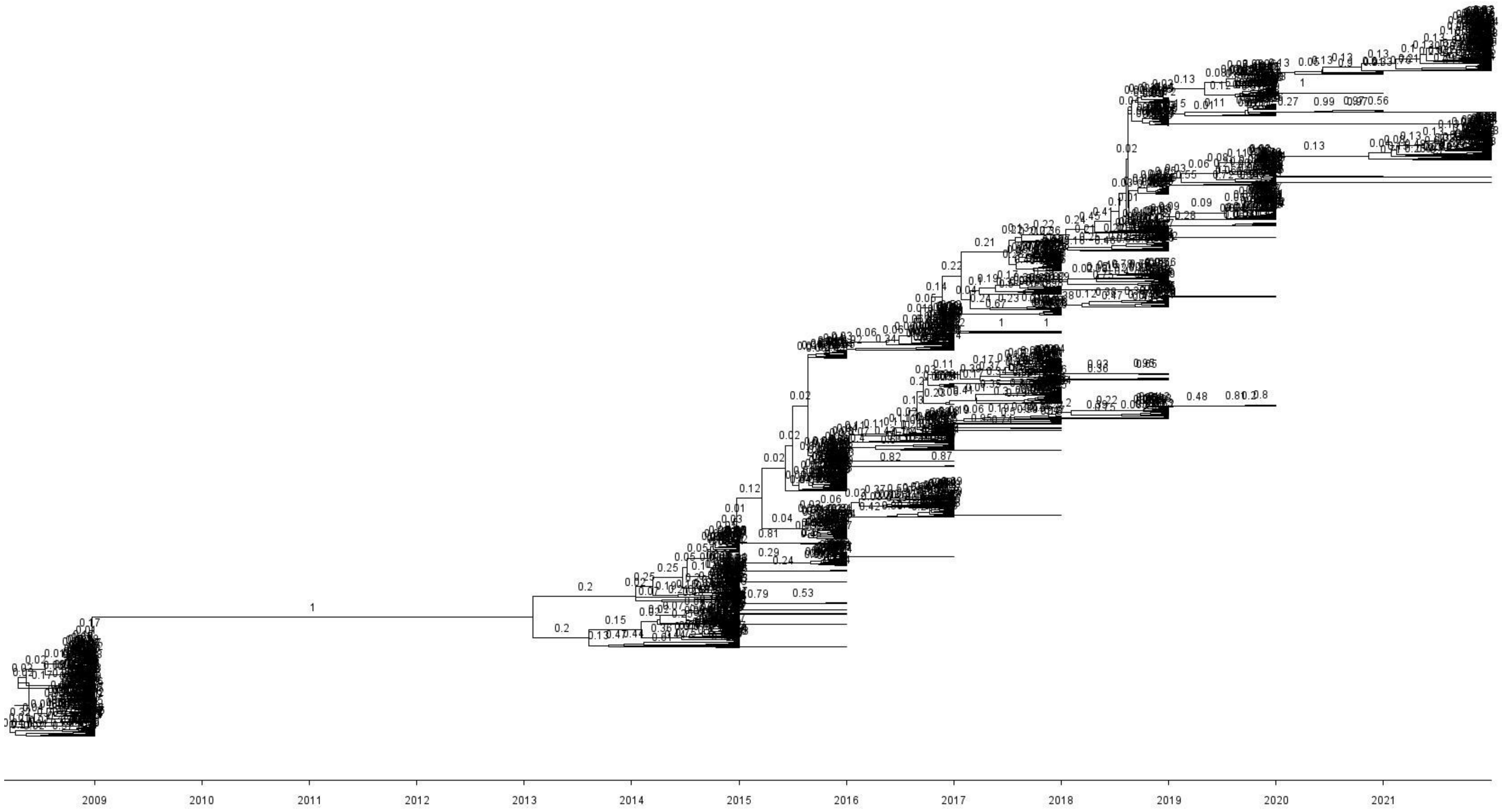
Supplementary Figure 40. Europe NA timetree 5 with posterior probabilities



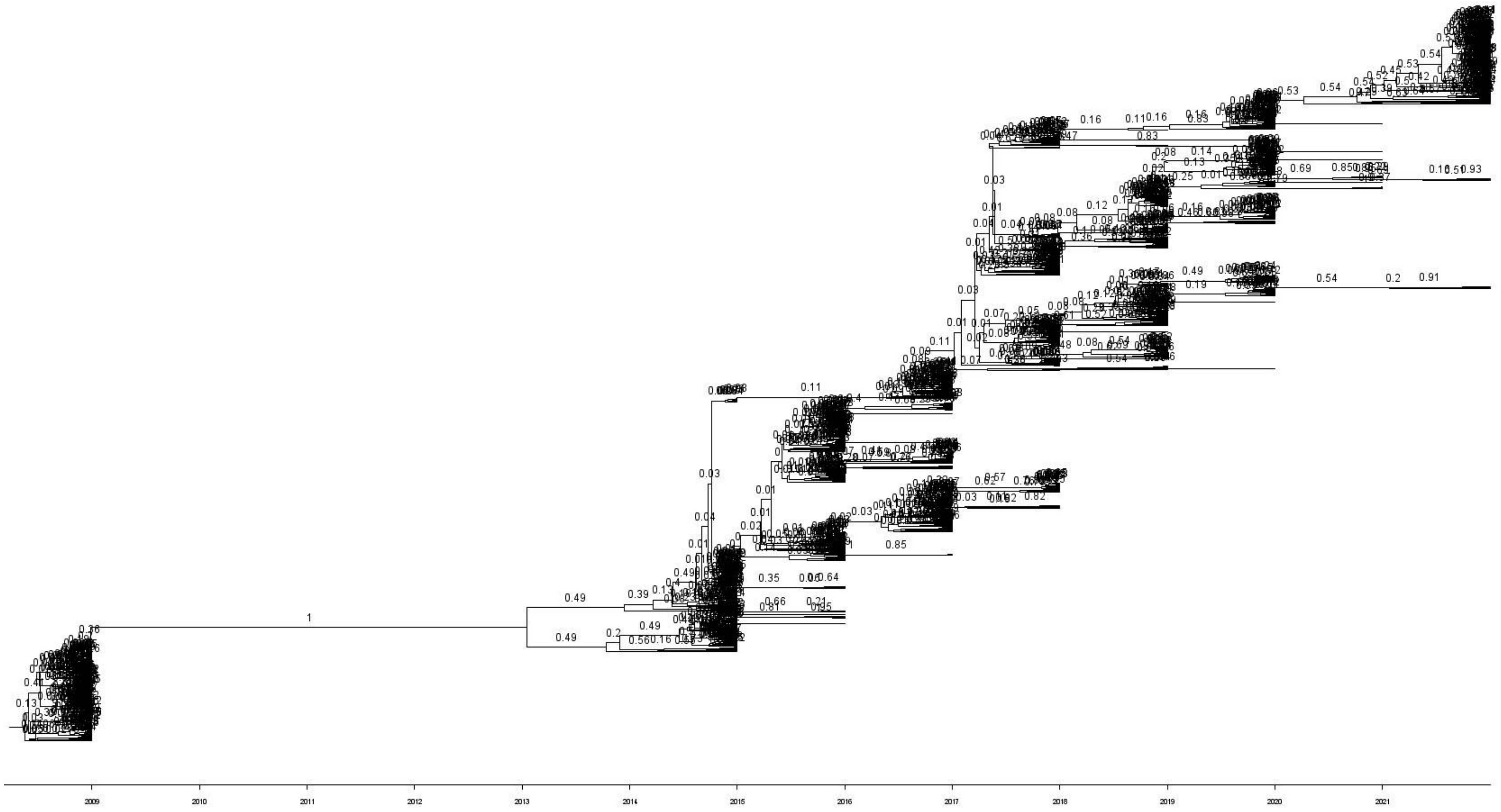
Supplementary Figure 41. North America HA timetree 1 with posterior probabilities



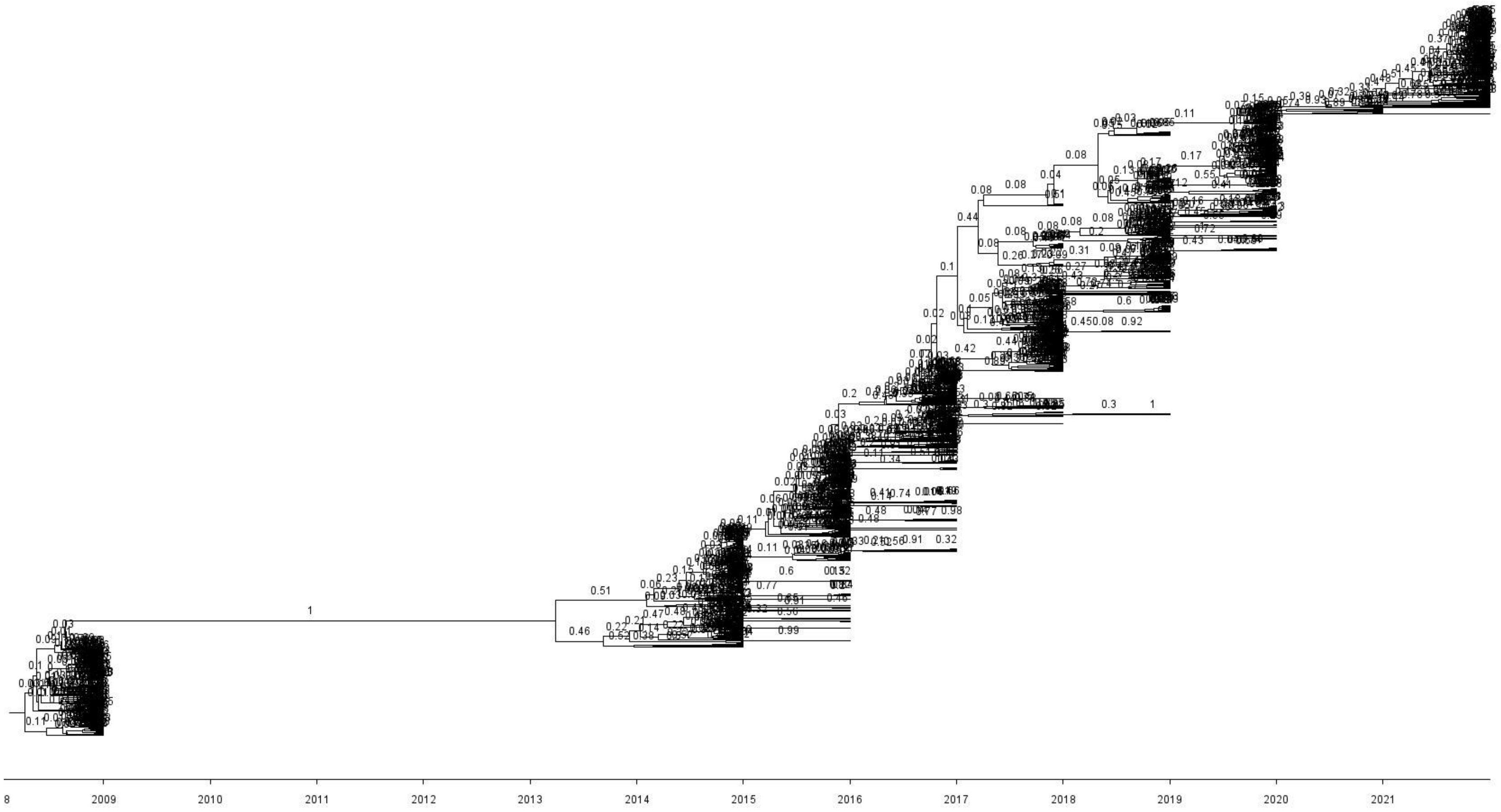
Supplementary Figure 42. North America HA timetree 2 with posterior probabilities



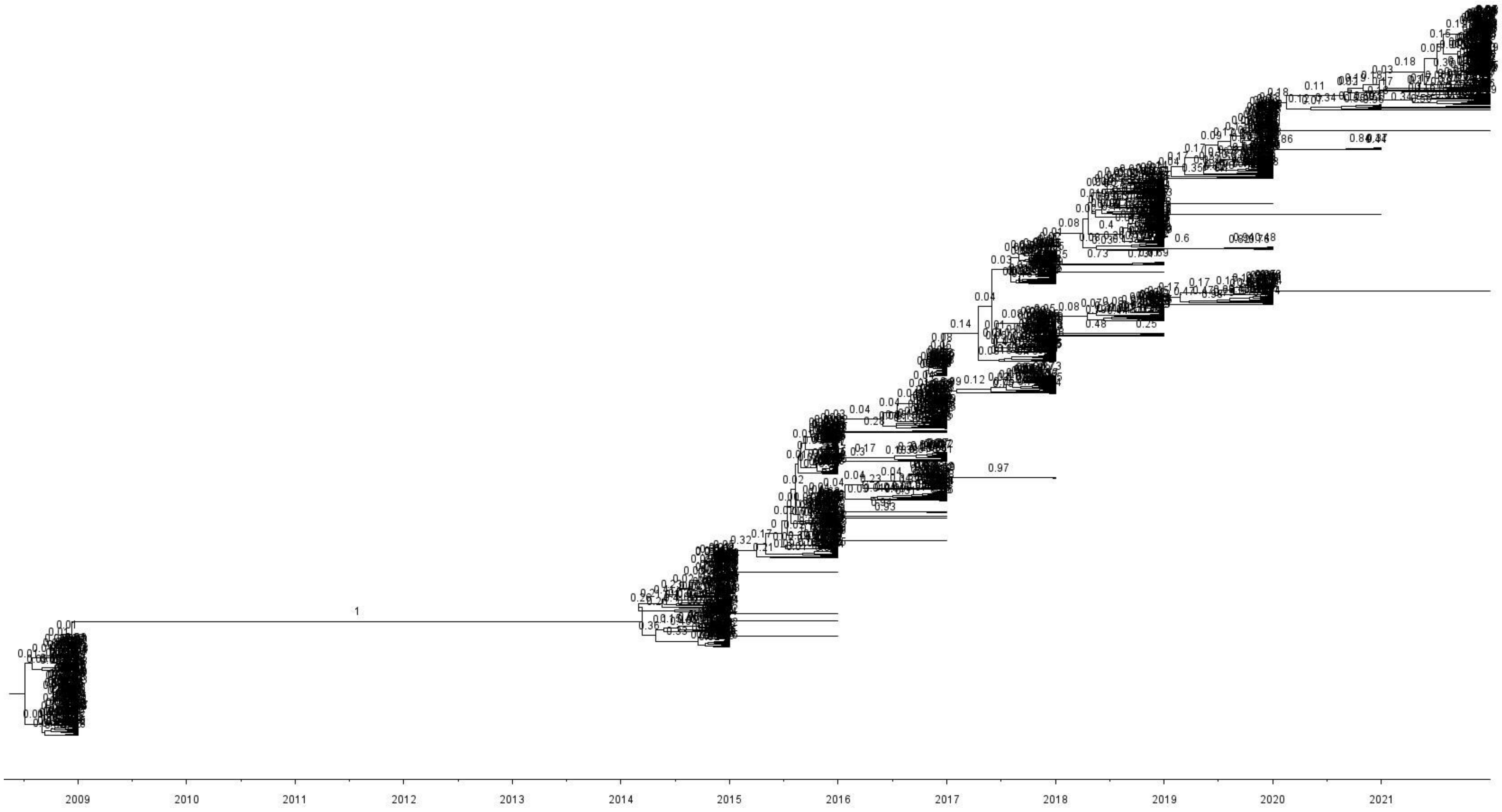
Supplementary Figure 43. North America HA timetree 3 with posterior probabilities



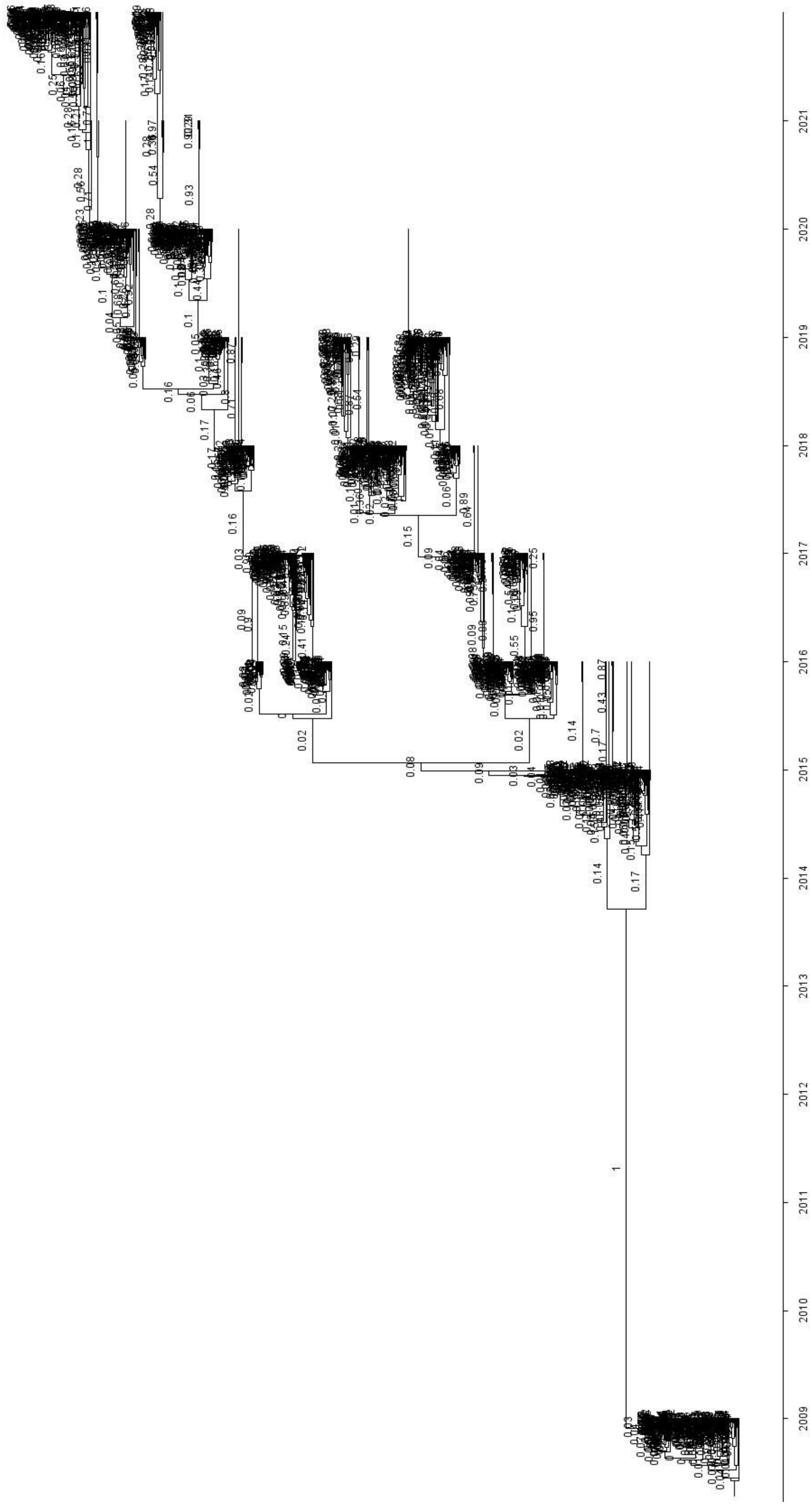
Supplementary Figure 44. North America HA timetree 4 with posterior probabilities



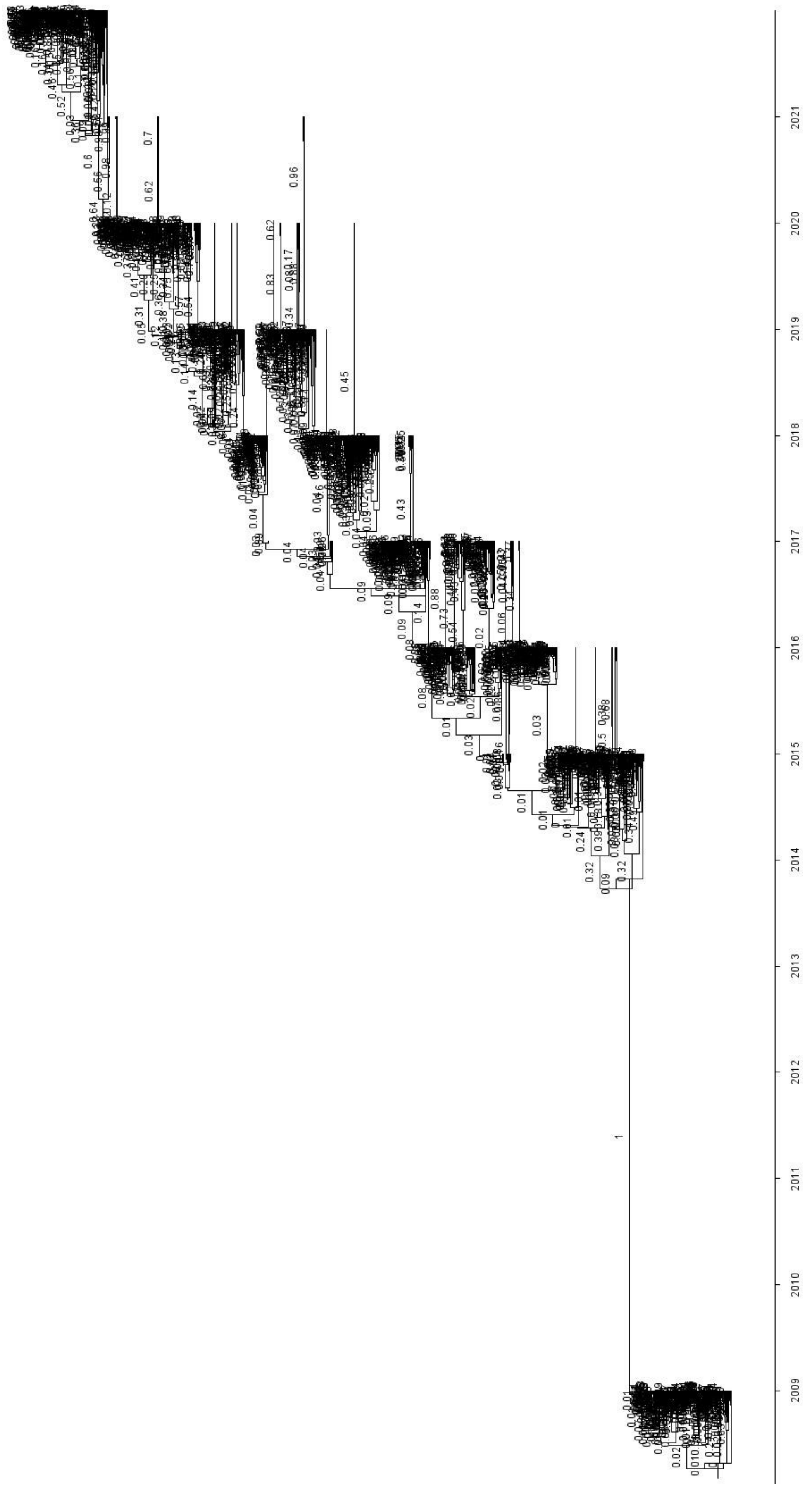
Supplementary Figure 45. North America HA timetree 5 with posterior probabilities



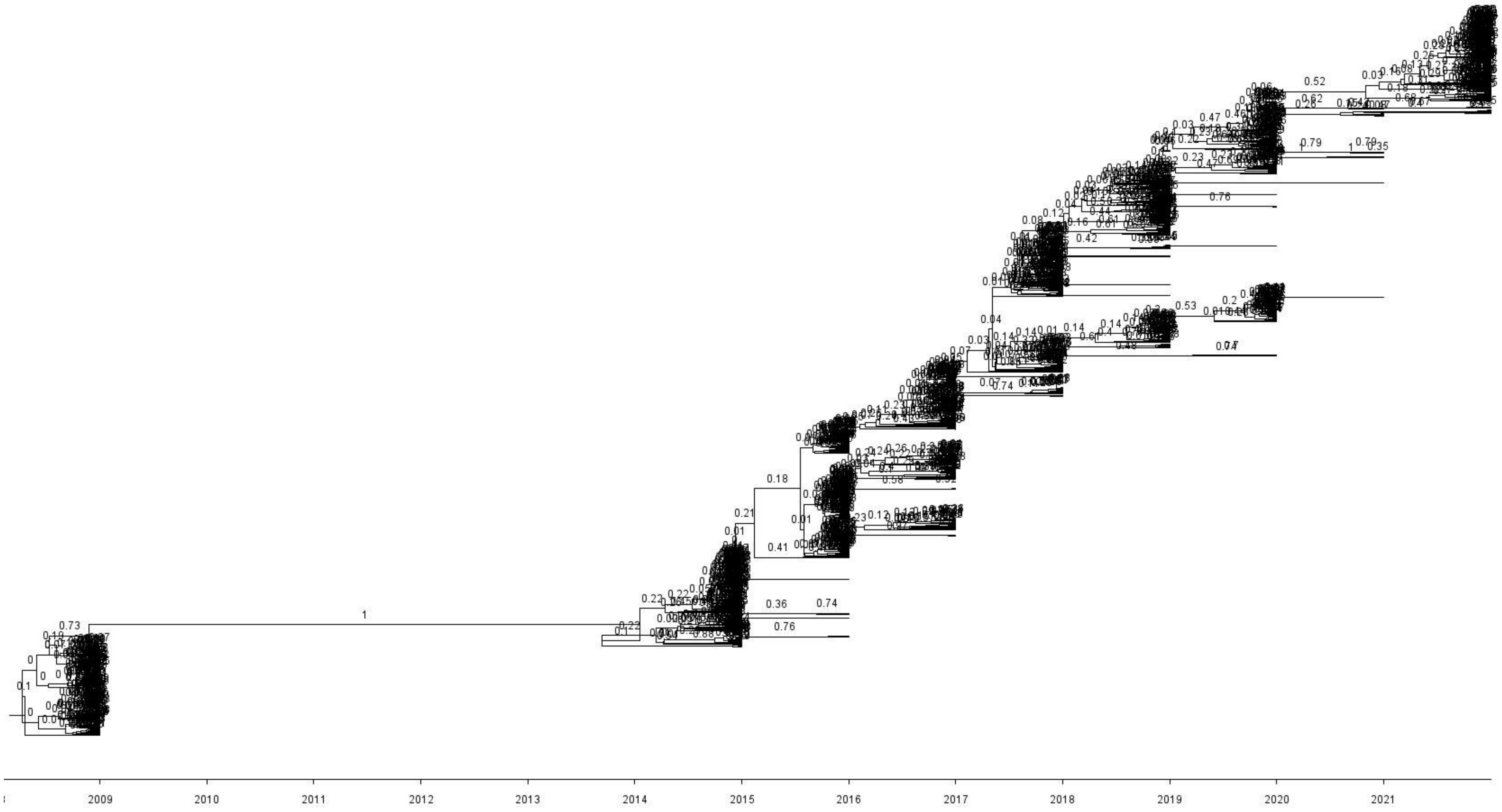
Supplementary Figure 46. North America NA timetree 1 with posterior probabilities



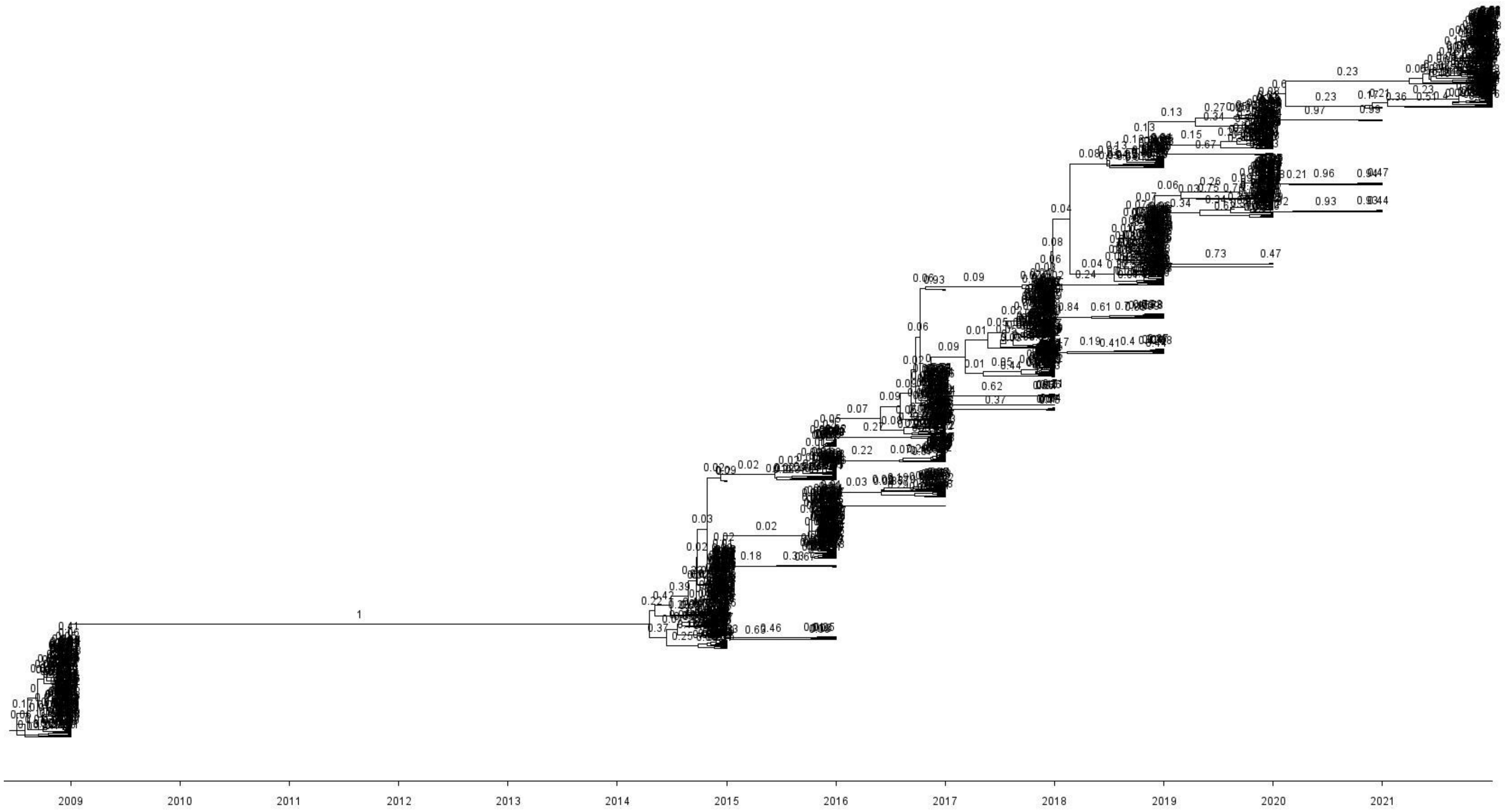
Supplementary Figure 47. North America NA timetree 2 with posterior probabilities



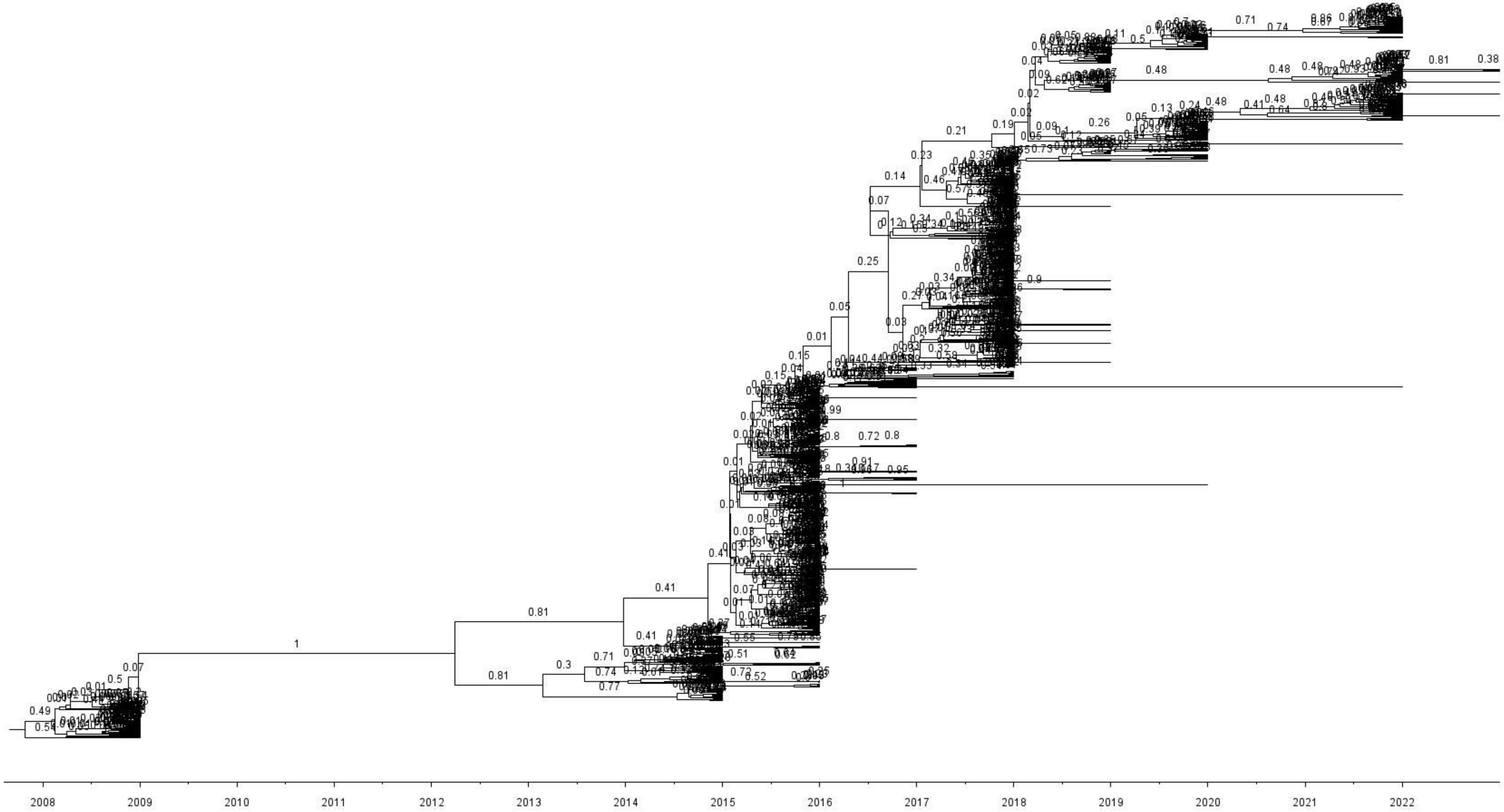
Supplementary Figure 48. North America NA timetree 3 with posterior probabilities



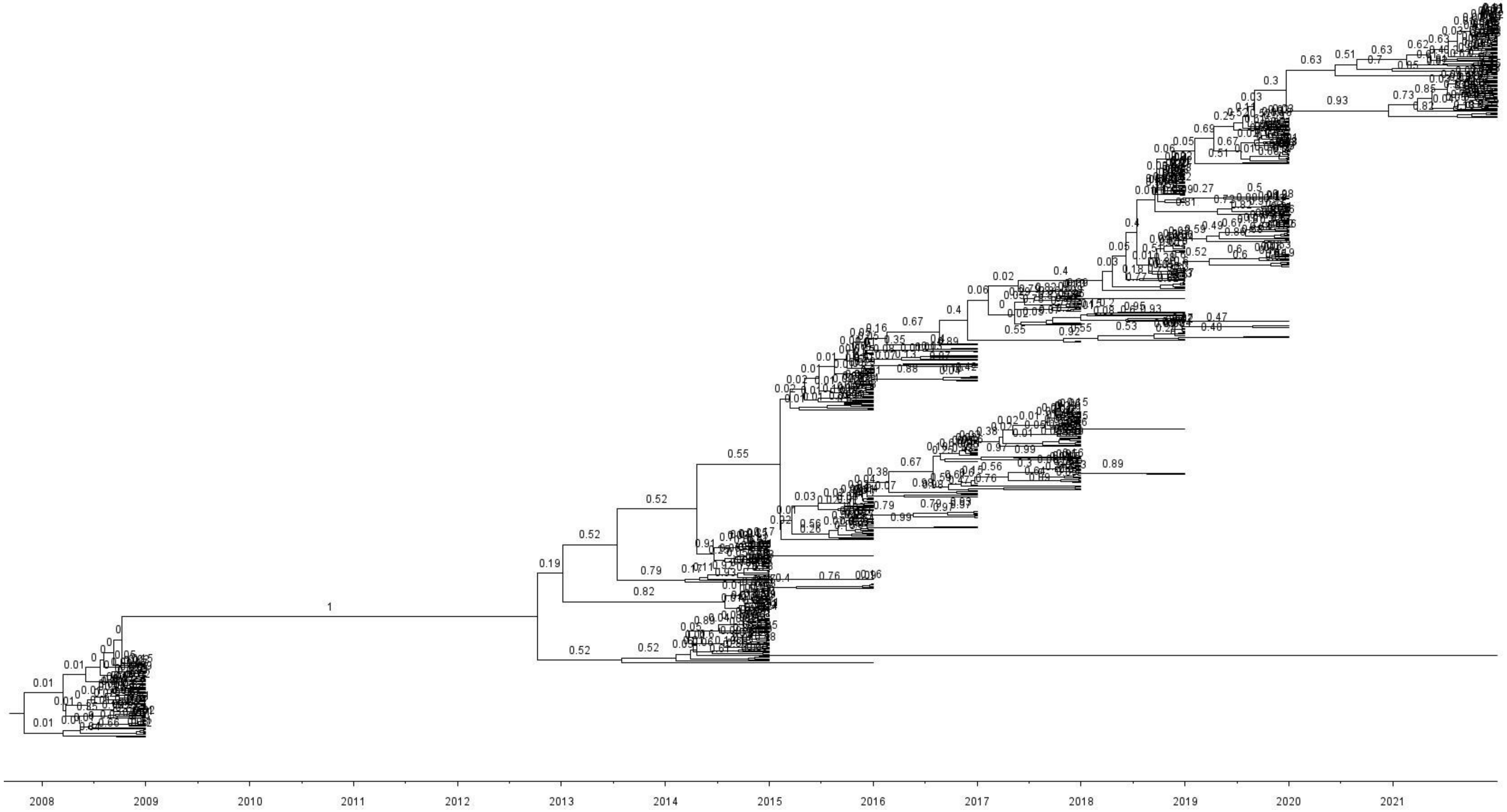
Supplementary Figure 49. North America NA timetree 4 with posterior probabilities



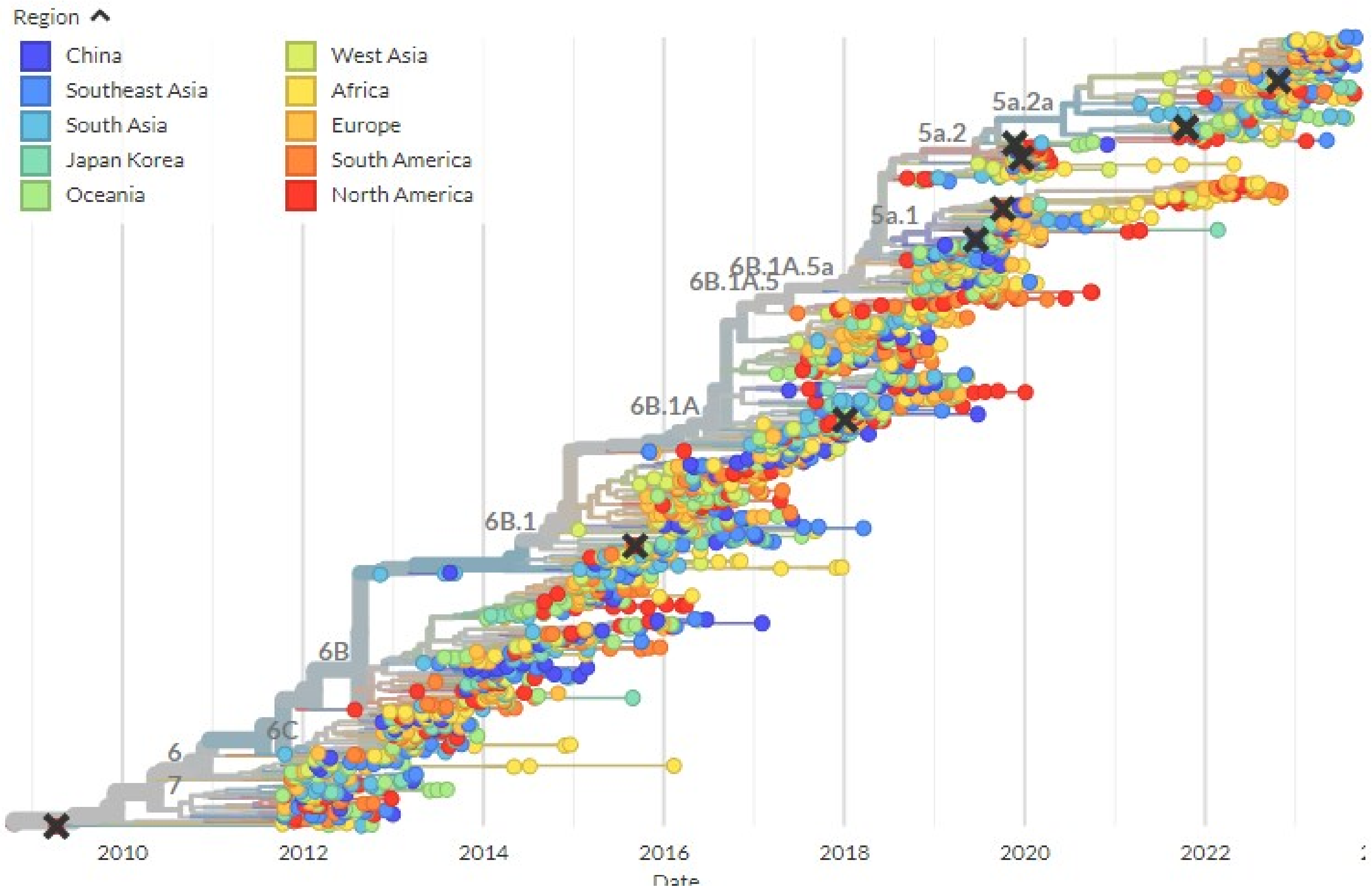
Supplementary Figure 50. North America NA timetree 5 with posterior probabilities



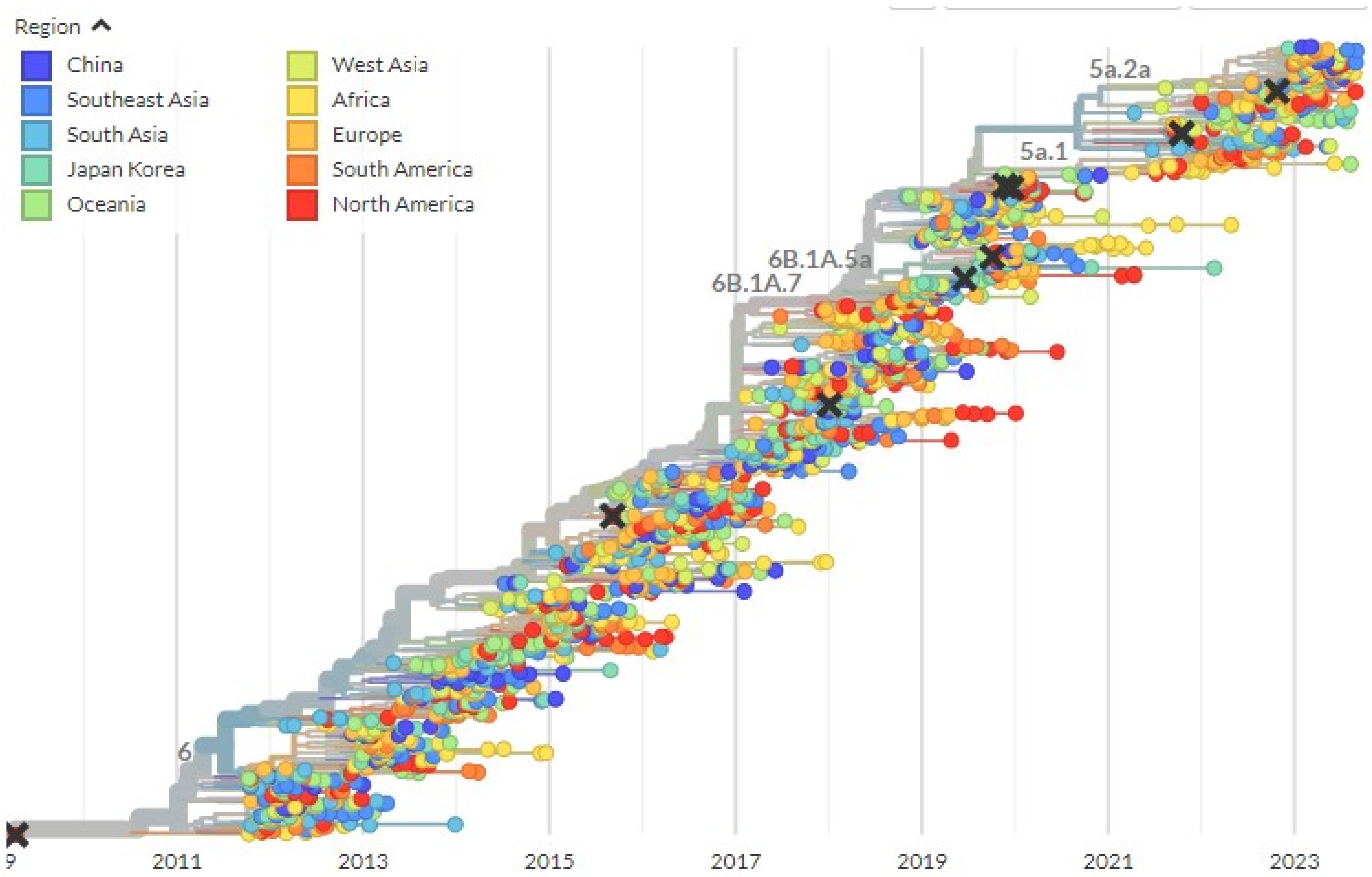
Supplementary Figure 51. South America HA timetree with posterior probabilities



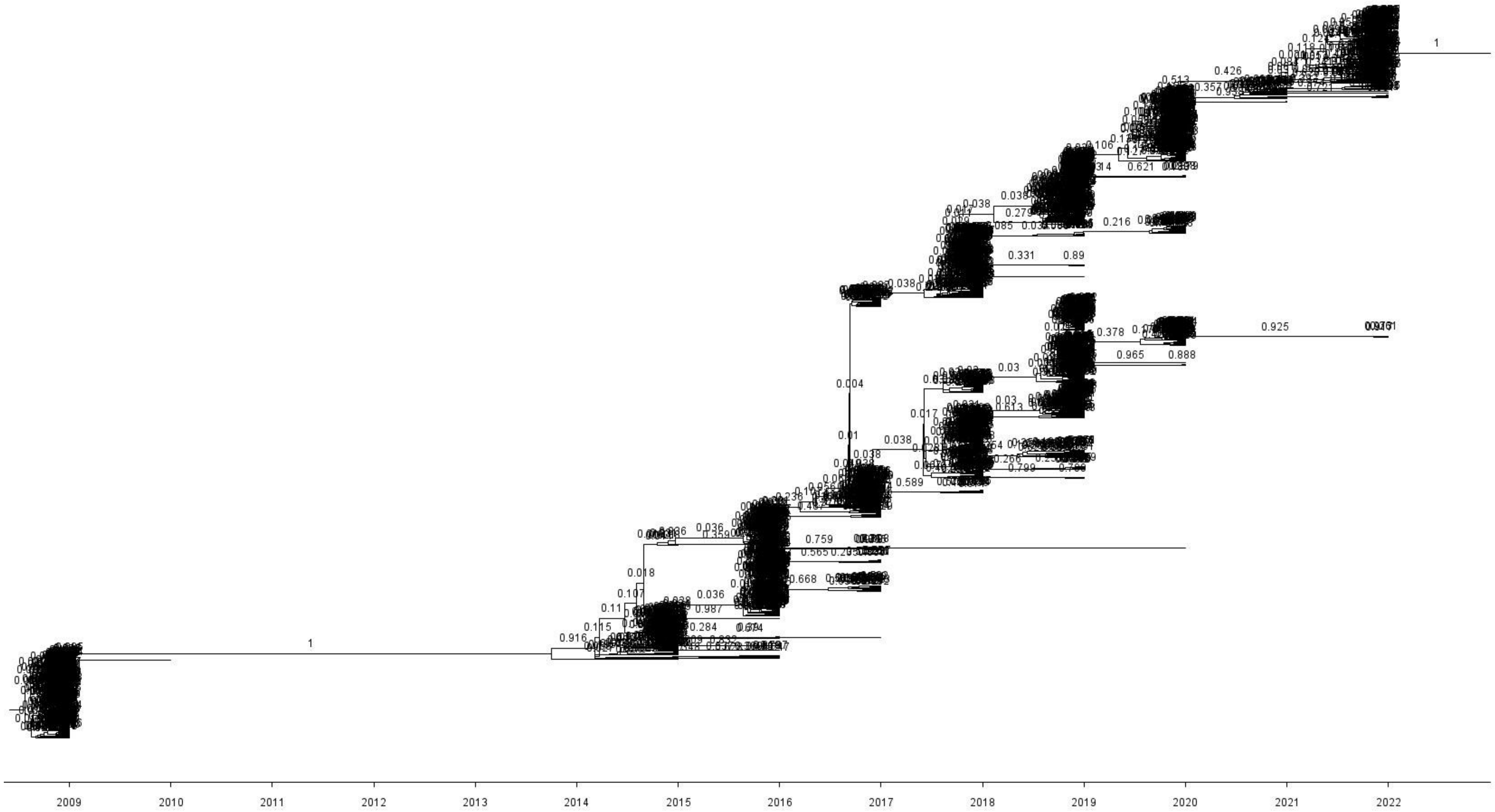
Supplementary Figure 52. South America NA timetree with posterior probabilities



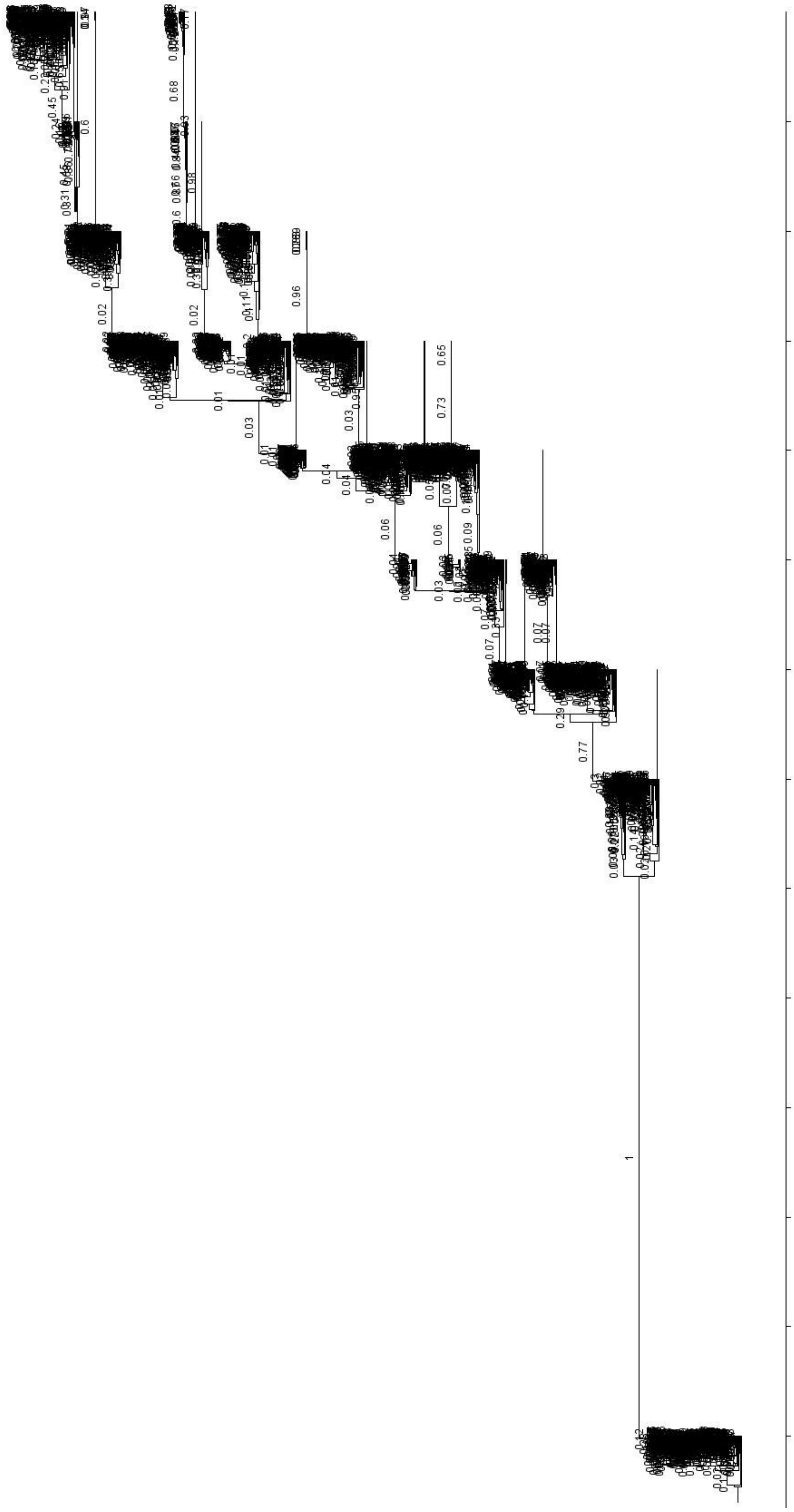
Supplementary Figure 53. GISAID phylogenetic timetree for HA sequences for 2009 pandemic H1N1.



Supplementary Figure 54. GISAID phylogenetic timetree for NA sequences for 2009 pandemic H1N1.

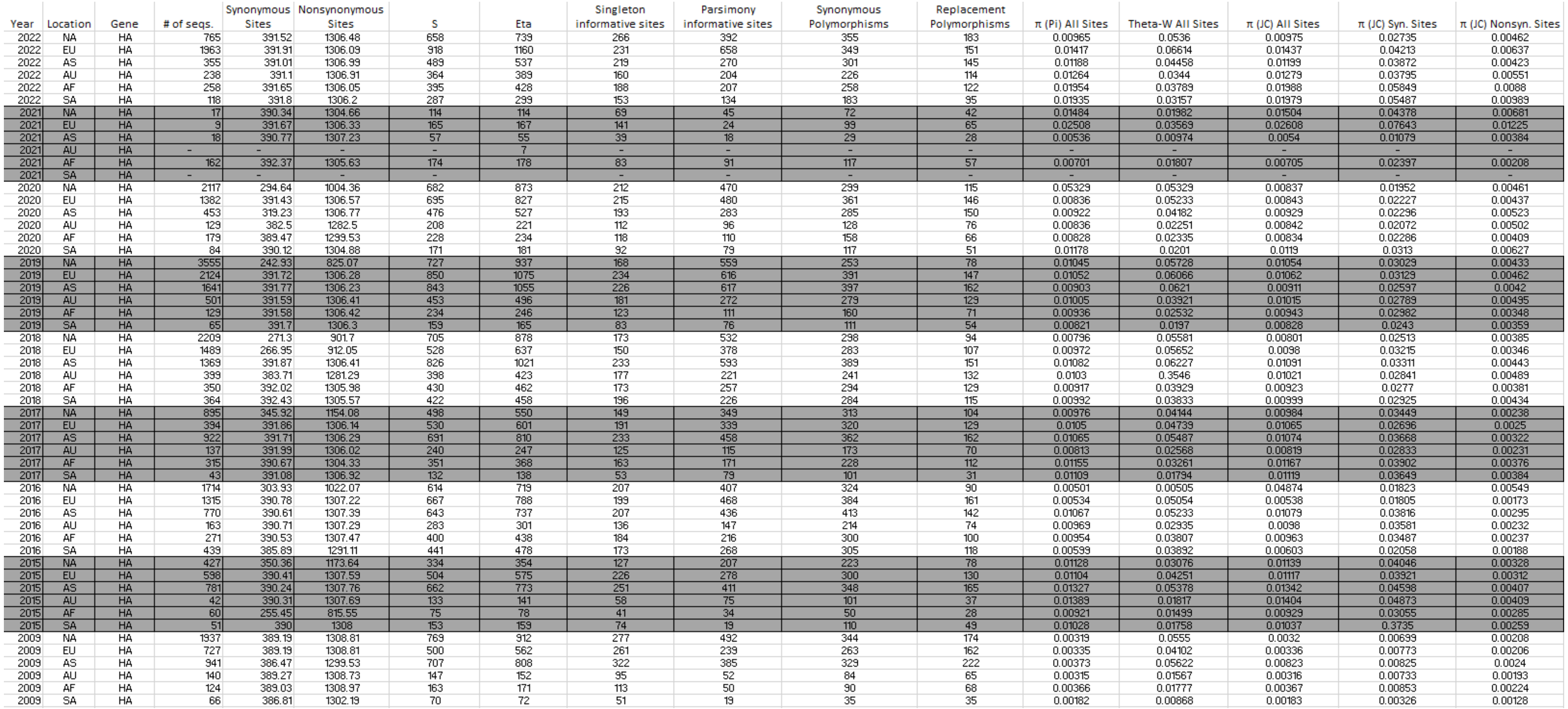


Supplementary Figure 55. Consensus HA tree with posterior probabilities



Supplementary Figure 56. Consensus NA tree with posterior probabilities

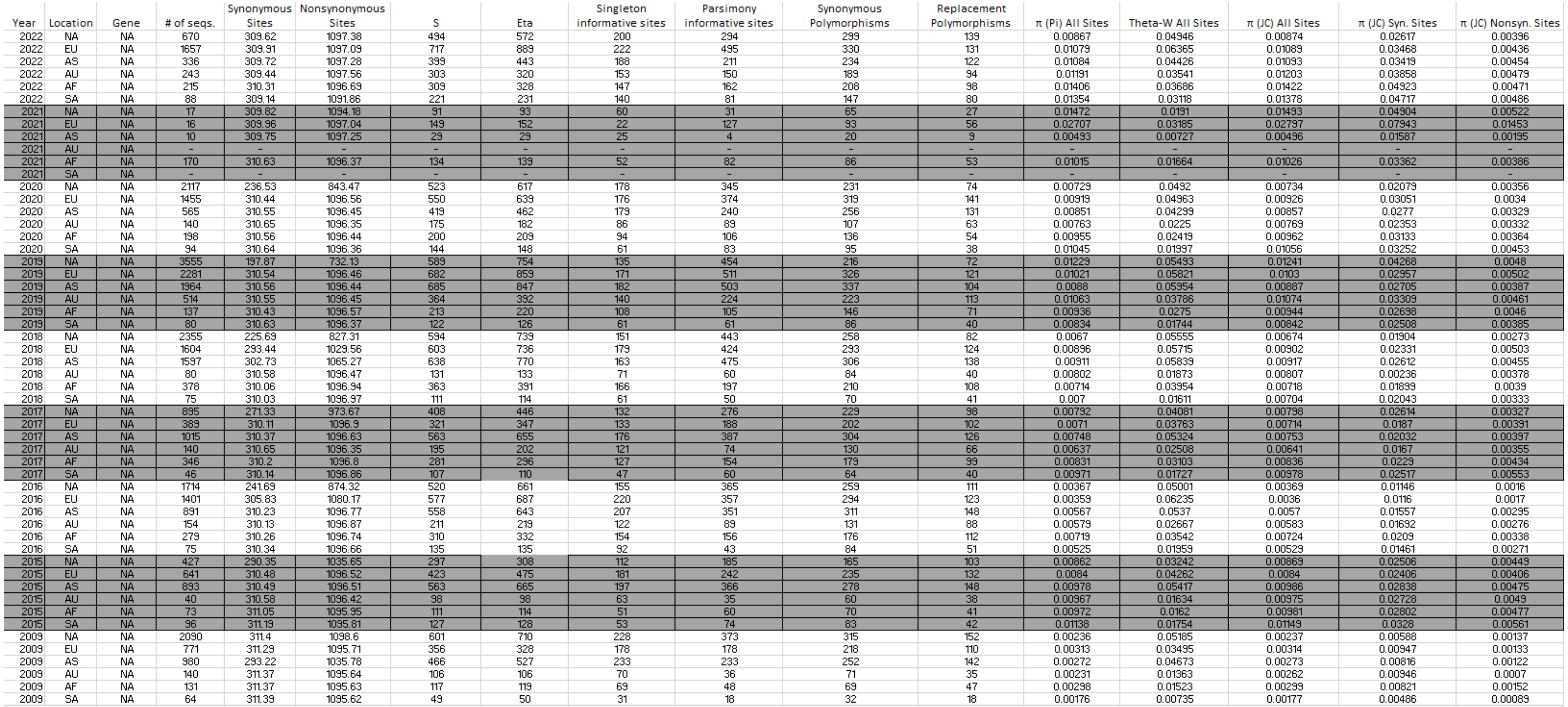
Supplemental Table 1. Population genetics analysis for nucleotide diversity for HA protein of 2009pdm influenza.



Year	Location	Gene	# of seqs.	Synonymous Sites	Nonsynonymous Sites	S	Eta	Singleton informative sites	Parsimony informative sites	Synonymous Polymorphisms	Replacement Polymorphisms	π (Pi) All Sites	Theta-W All Sites	π (JC) All Sites	π (JC) Syn. Sites	π (JC) Nonsyn. Sites
2022	NA	HA	765	391.52	1306.48	658	739	266	392	355	183	0.00965	0.0536	0.00975	0.02735	0.00462
2022	EU	HA	1963	391.91	1306.09	918	1160	231	658	349	151	0.01417	0.06614	0.01437	0.04213	0.00637
2022	AS	HA	355	391.01	1306.99	489	537	219	270	301	145	0.01188	0.04458	0.01199	0.03872	0.00423
2022	AU	HA	238	391.1	1306.91	364	389	160	204	226	114	0.01264	0.0344	0.01279	0.03795	0.00551
2022	AF	HA	258	391.65	1306.05	395	428	188	207	258	122	0.01954	0.03789	0.01988	0.05849	0.0088
2022	SA	HA	118	391.8	1306.2	287	299	153	134	183	95	0.01935	0.03157	0.01979	0.05487	0.00989
2021	NA	HA	17	390.34	1304.66	114	114	69	45	72	42	0.01484	0.01982	0.01504	0.04378	0.00681
2021	EU	HA	9	391.67	1306.33	165	167	141	24	99	65	0.02508	0.03569	0.02608	0.07643	0.01225
2021	AS	HA	18	390.77	1307.23	57	55	39	18	29	28	0.00536	0.00974	0.0054	0.01079	0.00384
2021	AU	HA	-	-	-	-	7	-	-	-	-	-	-	-	-	-
2021	AF	HA	162	392.37	1305.63	174	178	83	91	117	57	0.00701	0.01807	0.00705	0.02397	0.00208
2021	SA	HA	-	-	-	-		-	-	-	-	-	-	-	-	-
2020	NA	HA	2117	294.64	1004.36	682	873	212	470	299	115	0.05329	0.05329	0.00837	0.01952	0.00461
2020	EU	HA	1382	391.43	1306.57	695	827	215	480	361	146	0.00836	0.05233	0.00843	0.02227	0.00437
2020	AS	HA	453	319.23	1306.77	476	527	193	283	285	150	0.00922	0.04182	0.00929	0.02296	0.00523
2020	AU	HA	129	382.5	1282.5	208	221	112	96	128	76	0.00836	0.02251	0.00842	0.02072	0.00502
2020	AF	HA	179	389.47	1299.53	228	234	118	110	158	66	0.00828	0.02335	0.00834	0.02286	0.00409
2020	SA	HA	84	390.12	1304.88	171	181	92	79	117	51	0.01178	0.0201	0.0119	0.0313	0.00627
2019	NA	HA	3555	242.93	825.07	727	937	168	559	253	78	0.01045	0.05728	0.01054	0.03029	0.00433
2019	EU	HA	2124	391.72	1306.28	850	1075	234	616	391	147	0.01052	0.06066	0.01062	0.03129	0.00462
2019	AS	HA	1641	391.77	1306.23	843	1055	226	617	397	162	0.00903	0.0621	0.00911	0.02597	0.0042
2019	AU	HA	501	391.59	1306.41	453	496	181	272	279	129	0.01005	0.03921	0.01015	0.02789	0.00495
2019	AF	HA	129	391.58	1306.42	234	246	123	111	160	71	0.00936	0.02532	0.00943	0.02982	0.00348
2019	SA	HA	65	391.7	1306.3	159	165	83	76	111	54	0.00821	0.0197	0.00828	0.0243	0.00359
2018	NA	HA	2209	271.3	901.7	705	878	173	532	298	94	0.00796	0.05581	0.00801	0.02513	0.00385
2018	EU	HA	1489	266.95	912.05	528	637	150	378	283	107	0.00972	0.05652	0.0098	0.03215	0.00346
2018	AS	HA	1369	391.87	1306.41	826	1021	233	593	389	151	0.01082	0.06227	0.01091	0.03311	0.00443
2018	AU	HA	399	383.71	1281.29	398	423	177	221	241	132	0.0103	0.3546	0.01021	0.02841	0.00489
2018	AF	HA	350	392.02	1305.98	430	462	173	257	294	129	0.00917	0.03929	0.00923	0.0277	0.00381
2018	SA	HA	364	392.43	1305.57	422	458	196	226	284	115	0.00992	0.03833	0.00999	0.02925	0.00434
2017	NA	HA	895	345.92	1154.08	498	550	149	349	313	104	0.00976	0.04144	0.00984	0.03449	0.00238
2017	EU	HA	394	391.86	1306.14	530	601	191	339	320	129	0.0105	0.04739	0.01065	0.02696	0.0025
2017	AS	HA	922	391.71	1306.29	691	810	233	458	362	162	0.01065	0.05487	0.01074	0.03668	0.00322
2017	AU	HA	137	391.99	1306.02	240	247	125	115	173	70	0.00813	0.02568	0.00819	0.02833	0.00231
2017	AF	HA	315	390.67	1304.33	351	368	163	171	228	112	0.01155	0.03261	0.01167	0.03902	0.00376
2017	SA	HA	43	391.08	1306.92	132	138	53	79	101	31	0.01109	0.01794	0.01119	0.03649	0.00384
2016	NA	HA	1714	303.93	1022.07	614	719	207	407	324	90	0.00501	0.00505	0.04874	0.01823	0.00549
2016	EU	HA	1315	390.78	1307.22	667	788	199	468	384	161	0.00534	0.05054	0.00538	0.01805	0.00173
2016	AS	HA	770	390.61	1307.39	643	737	207	436	413	142	0.01067	0.05233	0.01079	0.03816	0.00295
2016	AU	HA	163	390.71	1307.29	283	301	136	147	214	74	0.00969	0.02935	0.0098	0.03581	0.00232
2016	AF	HA	271	390.53	1307.47	400	438	184	216	300	100	0.00954	0.03807	0.00963	0.03487	0.00237
2016	SA	HA	439	385.89	1291.11	441	478	173	268	305	118	0.00599	0.03892	0.00603	0.02058	0.00188
2015	NA	HA	427	350.36	1173.64	334	354	127	207	223	78	0.01128	0.03076	0.01139	0.04046	0.00328
2015	EU	HA	598	390.41	1307.59	504	575	226	278	300	130	0.01104	0.04251	0.01117	0.03921	0.00312
2015	AS	HA	781	390.24	1307.76	662	773	251	411	348	165	0.01327	0.05378	0.01342	0.04598	0.00407
2015	AU	HA	42	390.31	1307.69	133	141	58	75	101	37	0.01389	0.01817	0.01404	0.04873	0.00409
2015	AF	HA	60	255.45	815.55	75	78	41	34	50	28	0.00921	0.01499	0.00929	0.03055	0.00285
2015	SA	HA	51	390	1308	153	159	74	19	110	49	0.01028	0.01758	0.01037	0.3735	0.00259
2009	NA	HA	1937	389.19	1308.81	769	912	277	492	344	174	0.00319	0.0555	0.0032	0.00699	0.00208
2009	EU	HA	727	389.19	1308.81	500	562	261	239	263	162	0.00335	0.04102	0.00336	0.00773	0.00206
2009	AS	HA	941	386.47	1299.53	707	808	322	385	329	222	0.00373	0.05622	0.00823	0.00825	0.0024
2009	AU	HA	140	389.27	1308.73	147	152	95	52	84	65	0.00315	0.01567	0.00316	0.00733	0.00193
2009	AF	HA	124	389.03	1308.97	163	171	113	50	90	68	0.00366	0.01777	0.00367	0.00853	0.00224
2009	SA	HA	66	386.81	1302.19	70	72	51	19	35	35	0.00182	0.00868	0.00183	0.00326	0.00128

(S = segregating sites, Eta = total number of mutations, π = nucleotide diversity, JC = Jukes-Cantor corrected)

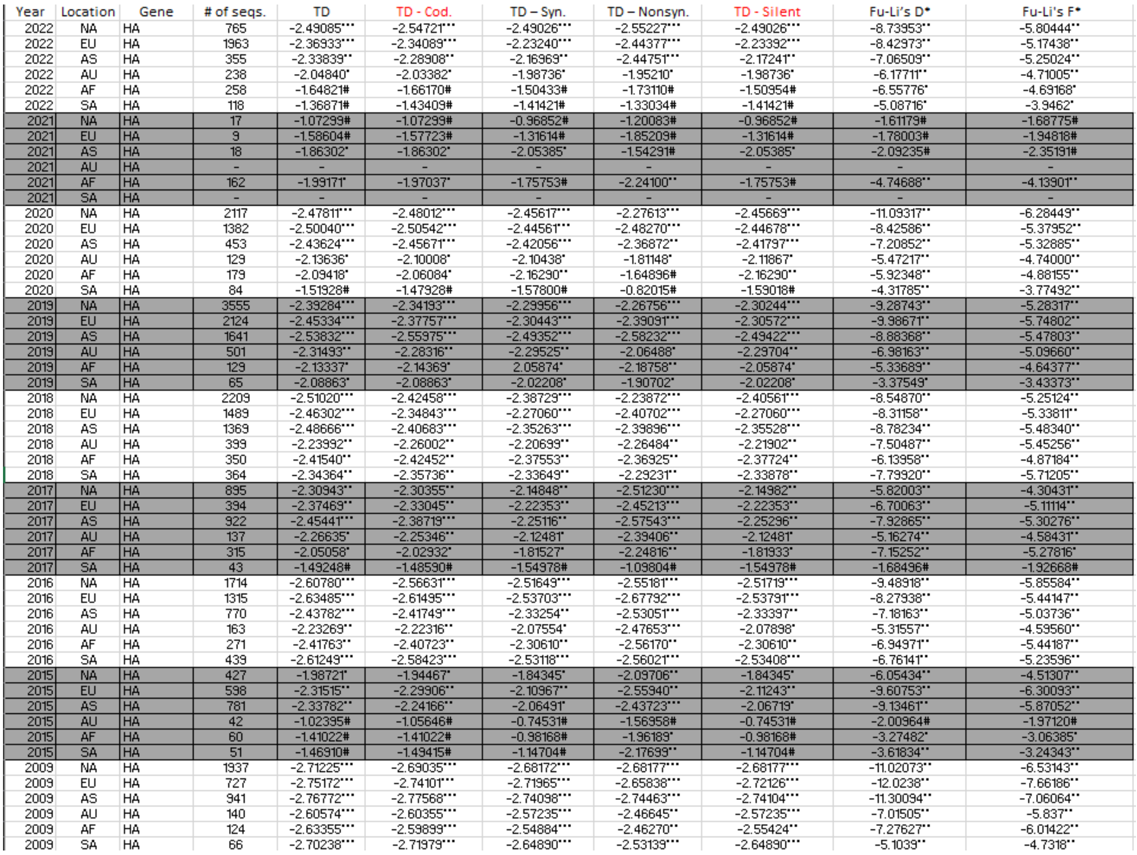
Supplementary Table 2. Population genetics analysis for nucleotide diversity for NA protein of 2009pdm influenza.



Year	Location	Gene	# of seqs.	Synonymous Sites	Nonsynonymous Sites	S	Eta	Singleton informative sites	Parsimony informative sites	Synonymous Polymorphisms	Replacement Polymorphisms	π (Pi) All Sites	Theta-W All Sites	π (JC) All Sites	π (JC) Syn. Sites	π (JC) Nonsyn. Sites
2022	NA	NA	670	309.62	1097.38	494	572	200	294	299	139	0.00867	0.04946	0.00874	0.02617	0.00396
2022	EU	NA	1657	309.91	1097.09	717	889	222	495	330	131	0.01079	0.06365	0.01089	0.03468	0.00436
2022	AS	NA	336	309.72	1097.28	399	443	188	211	234	122	0.01084	0.04426	0.01093	0.03419	0.00454
2022	AU	NA	243	309.44	1097.56	303	320	153	150	189	94	0.01191	0.03541	0.01203	0.03858	0.00479
2022	AF	NA	215	310.31	1096.69	309	328	147	162	208	98	0.01406	0.03686	0.01422	0.04923	0.00471
2022	SA	NA	88	309.14	1091.86	221	231	140	81	147	80	0.01354	0.03118	0.01378	0.04717	0.00486
2021	NA	NA	17	309.82	1094.18	91	93	60	31	65	27	0.01472	0.0191	0.01493	0.04904	0.00522
2021	EU	NA	16	309.96	1097.04	149	152	22	127	93	56	0.02707	0.03185	0.02797	0.07943	0.01453
2021	AS	NA	10	309.75	1097.25	29	29	25	4	20	9	0.00493	0.00727	0.00496	0.01587	0.00195
2021	AU	NA	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2021	AF	NA	170	310.63	1096.37	134	139	52	82	86	53	0.01015	0.01664	0.01026	0.03362	0.00386
2021	SA	NA	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2020	NA	NA	2117	236.53	843.47	523	617	178	345	231	74	0.00729	0.0492	0.00734	0.02079	0.00356
2020	EU	NA	1455	310.44	1096.56	550	639	176	374	319	141	0.00919	0.04963	0.00926	0.03051	0.0034
2020	AS	NA	565	310.55	1096.45	419	462	179	240	256	131	0.00851	0.04299	0.00857	0.0277	0.00329
2020	AU	NA	140	310.65	1096.35	175	182	86	89	107	63	0.00763	0.0225	0.00769	0.02353	0.00332
2020	AF	NA	198	310.56	1096.44	200	209	94	106	136	54	0.00955	0.02419	0.00962	0.03133	0.00364
2020	SA	NA	94	310.64	1096.36	144	148	61	83	95	38	0.01045	0.01997	0.01056	0.03252	0.00453
2019	NA	NA	3555	197.87	732.13	589	754	135	454	216	72	0.01229	0.05493	0.01241	0.04268	0.0048
2019	EU	NA	2281	310.54	1096.46	682	859	171	511	326	121	0.01021	0.05821	0.0103	0.02957	0.00502
2019	AS	NA	1964	310.56	1096.44	685	847	182	503	337	104	0.0088	0.05954	0.00887	0.02705	0.00387
2019	AU	NA	514	310.55	1096.45	364	392	140	224	223	113	0.01063	0.03786	0.01074	0.03309	0.00461
2019	AF	NA	137	310.43	1096.57	213	220	108	105	146	71	0.00936	0.0275	0.00944	0.02698	0.0046
2019	SA	NA	80	310.63	1096.37	122	126	61	61	86	40	0.00834	0.01744	0.00842	0.02508	0.00385
2018	NA	NA	2355	225.69	827.31	594	739	151	443	258	82	0.0067	0.05555	0.00674	0.01904	0.00273
2018	EU	NA	1604	293.44	1029.56	603	736	179	424	293	124	0.00896	0.05715	0.00902	0.02331	0.00503
2018	AS	NA	1597	302.73	1065.27	638	770	163	475	306	138	0.00911	0.05839	0.00917	0.02612	0.00455
2018	AU	NA	80	310.58	1096.47	131	133	71	60	84	40	0.00802	0.01873	0.00807	0.00236	0.00378
2018	AF	NA	378	310.06	1096.94	363	391	166	197	210	108	0.00714	0.03954	0.00718	0.01899	0.0039
2018	SA	NA	75	310.03	1096.97	111	114	61	50	70	41	0.007	0.01611	0.00704	0.02043	0.00333
2017	NA	NA	895	271.33	973.67	408	446	132	276	229	98	0.00792	0.04081	0.00798	0.02614	0.00327
2017	EU	NA	389	310.11	1096.9	321	347	133	188	202	102	0.0071	0.03763	0.00714	0.0187	0.00391
2017	AS	NA	1015	310.37	1096.63	563	655	176	387	304	126	0.00748	0.05324	0.00753	0.02032	0.00397
2017	AU	NA	140	310.65	1096.35	195	202	121	74	130	66	0.00637	0.02508	0.00641	0.0167	0.00355
2017	AF	NA	346	310.2	1096.8	281	296	127	154	179	99	0.00831	0.03103	0.00836	0.0229	0.00434
2017	SA	NA	46	310.14	1096.86	107	110	47	60	64	40	0.00971	0.01727	0.00978	0.02517	0.00553
2016	NA	NA	1714	241.69	874.32	520	661	155	365	259	111	0.00367	0.05001	0.00369	0.01146	0.0016
2016	EU	NA	1401	305.83	1080.17	577	687	220	357	294	123	0.00359	0.06235	0.0036	0.0116	0.0017
2016	AS	NA	891	310.23	1096.77	558	643	207	351	311	148	0.00567	0.0537	0.0057	0.01557	0.00295
2016	AU	NA	154	310.13	1096.87	211	219	122	89	131	88	0.00579	0.02667	0.00583	0.01692	0.00276
2016	AF	NA	279	310.26	1096.74	310	332	154	156	176	112	0.00719	0.03542	0.00724	0.0209	0.00338
2016	SA	NA	75	310.34	1096.66	135	135	92	43	84	51	0.00525	0.01959	0.00529	0.01461	0.00271
2015	NA	NA	427	290.35	1035.65	297	308	112	185	165	103	0.00862	0.03242	0.00869	0.02506	0.00449
2015	EU	NA	641	310.48	1096.52	423	475	181	242	235	132	0.0084	0.04262	0.0084	0.02406	0.00406
2015	AS	NA	893	310.49	1096.51	563	665	197	366	278	148	0.00978	0.05417	0.00986	0.02838	0.00475
2015	AU	NA	40	310.58	1096.42	98	98	63	35	60	38	0.00967	0.01634	0.00975	0.02728	0.0049
2015	AF	NA	73	311.05	1095.95	111	114	51	60	70	41	0.00972	0.0162	0.00981	0.02802	0.00477
2015	SA	NA	96	311.19	1095.81	127	128	53	74	83	42	0.01138	0.01754	0.01149	0.0328	0.00561
2009	NA	NA	2090	311.4	1098.6	601	710	228	373	315	152	0.00236	0.05185	0.00237	0.00588	0.00137
2009	EU	NA	771	311.29	1095.71	356	328	178	178	218	110	0.00313	0.03495	0.00314	0.00947	0.00133
2009	AS	NA	980	293.22	1035.78	466	527	233	233	252	142	0.00272	0.04673	0.00273	0.00816	0.00122
2009	AU	NA	140	311.37	1095.64	106	106	70	36	71	35	0.00231	0.01363	0.00262	0.00946	0.0007
2009	AF	NA	131	311.37	1095.63	117	119	69	48	69	47	0.00298	0.01523	0.00299	0.00821	0.00152
2009	SA	NA	64	311.39	1095.62	49	50	31	18	32	18	0.00176	0.00735	0.00177	0.00486	0.00089

(S = segregating sites, Eta = total number of mutations, π = nucleotide diversity, JC = Jukes-Cantor corrected)

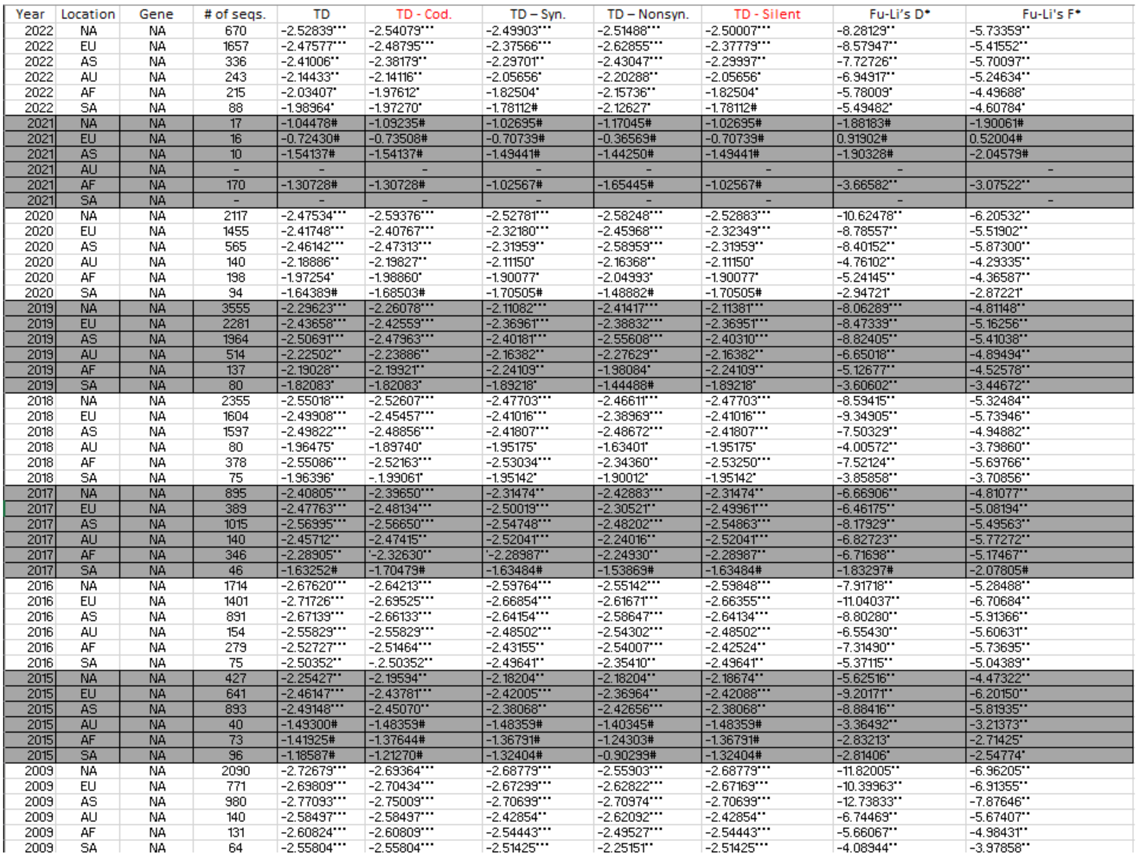
Supplementary Table 3. Tajima’s D and Fu-Li’s D* and F* tests for neutrality selection for HA protein of 2009pdm influenza.



Year	Location	Gene	# of seqs.	TD	TD - Cod.	TD – Syn.	TD – Nonsyn.	TD - Silent	Fu-Li’s D*	Fu-Li's F*
2022	NA	HA	765	-2.49085***	-2.54721***	-2.49026***	-2.55227***	-2.49026***	-8.73953**	-5.80444**
2022	EU	HA	1963	-2.36933***	-2.34089***	-2.23240***	-2.44377***	-2.23392***	-8.42973**	-5.17438**
2022	AS	HA	355	-2.33839**	-2.28908**	-2.16969**	-2.44751***	-2.17241**	-7.06509**	-5.25024**
2022	AU	HA	238	-2.04840*	-2.03382*	-1.98736*	-1.95210*	-1.98736*	-6.17711**	-4.71005**
2022	AF	HA	258	-1.64821#	-1.66170#	-1.50433#	-1.73110#	-1.50954#	-6.55776*	-4.69168*
2022	SA	HA	118	-1.36871#	-1.43409#	-1.41421#	-1.33034#	-1.41421#	-5.08716*	-3.9462*
2021	NA	HA	17	-1.07299#	-1.07299#	-0.96852#	-1.20083#	-0.96852#	-1.61179#	-1.68775#
2021	EU	HA	9	-1.58604#	-1.57723#	-1.31614#	-1.85209#	-1.31614#	-1.78003#	-1.94818#
2021	AS	HA	18	-1.86302*	-1.86302*	-2.05385*	-1.54291#	-2.05385*	-2.09235#	-2.35191#
2021	AU	HA	-	-	-	-	-	-	-	-
2021	AF	HA	162	-1.99171*	-1.97037*	-1.75753#	-2.24100**	-1.75753#	-4.74688**	-4.13901**
2021	SA	HA	-	-	-	-	-	-	-	-
2020	NA	HA	2117	-2.47811***	-2.48012***	-2.45617***	-2.27613***	-2.45669***	-11.09317**	-6.28449**
2020	EU	HA	1382	-2.50040***	-2.50542***	-2.44561***	-2.48270***	-2.44678***	-8.42586**	-5.37952**
2020	AS	HA	453	-2.43624***	-2.45671***	-2.42056***	-2.36872**	-2.41797***	-7.20852**	-5.32885**
2020	AU	HA	129	-2.13636*	-2.10008*	-2.10438*	-1.81148*	-2.11867*	-5.47217**	-4.74000**
2020	AF	HA	179	-2.09418*	-2.06084*	-2.16290**	-1.64896#	-2.16290**	-5.92348**	-4.88155**
2020	SA	HA	84	-1.51928#	-1.47928#	-1.57800#	-0.82015#	-1.59018#	-4.31785**	-3.77492**
2019	NA	HA	3555	-2.39284***	-2.34193***	-2.29956***	-2.26756***	-2.30244***	-9.28743**	-5.28317**
2019	EU	HA	2124	-2.45334***	-2.37757***	-2.30443***	-2.39091***	-2.30572***	-9.98671**	-5.74802**
2019	AS	HA	1641	-2.53832***	-2.55975***	-2.49352***	-2.58232***	-2.49422***	-8.88368**	-5.47803**
2019	AU	HA	501	-2.31493**	-2.28316**	-2.29525**	-2.06488*	-2.29704**	-6.98163**	-5.09660**
2019	AF	HA	129	-2.13337*	-2.14369*	2.05874*	-2.18758**	-2.05874*	-5.33689**	-4.64377**
2019	SA	HA	65	-2.08863*	-2.08863*	-2.02208*	-1.90702*	-2.02208*	-3.37549*	-3.43373**
2018	NA	HA	2209	-2.51020***	-2.42458***	-2.38729***	-2.23872***	-2.40561***	-8.54870**	-5.25124**
2018	EU	HA	1489	-2.46302***	-2.34843***	-2.27060***	-2.40702***	-2.27060***	-8.31158**	-5.33811**
2018	AS	HA	1369	-2.48666***	-2.40683***	-2.35263***	-2.39896***	-2.35528***	-8.78234**	-5.48340**
2018	AU	HA	399	-2.23992**	-2.26002**	-2.20699**	-2.26484**	-2.21902**	-7.50487**	-5.45256**
2018	AF	HA	350	-2.41540**	-2.42452**	-2.37553**	-2.36925**	-2.37724**	-6.13958**	-4.87184**
2018	SA	HA	364	-2.34364**	-2.35736**	-2.33649**	-2.29231**	-2.33878**	-7.79920**	-5.71205**
2017	NA	HA	895	-2.30943**	-2.30355**	-2.14848**	-2.51230***	-2.14982**	-5.82003**	-4.30431**
2017	EU	HA	394	-2.37469**	-2.33045**	-2.22353**	-2.45213***	-2.22353**	-6.70063**	-5.11114**
2017	AS	HA	922	-2.45441***	-2.38719***	-2.25116**	-2.57543***	-2.25296**	-7.92865**	-5.30276**
2017	AU	HA	137	-2.26635*	-2.25346**	-2.12481*	-2.39406**	-2.12481*	-5.16274**	-4.58431**
2017	AF	HA	315	-2.05058*	-2.02932*	-1.81527*	-2.24816**	-1.81933*	-7.15252**	-5.27816*
2017	SA	HA	43	-1.49248#	-1.48590#	-1.54978#	-1.09804#	-1.54978#	-1.68496#	-1.92668#
2016	NA	HA	1714	-2.60780***	-2.56631***	-2.51649***	-2.55181***	-2.51719***	-9.48918**	-5.85584**
2016	EU	HA	1315	-2.63485***	-2.61495***	-2.53703***	-2.67792***	-2.53791***	-8.27938**	-5.44147**
2016	AS	HA	770	-2.43782***	-2.41749***	-2.33254**	-2.53051***	-2.33397**	-7.18163**	-5.03736**
2016	AU	HA	163	-2.23269**	-2.22316**	-2.07554*	-2.47653***	-2.07898*	-5.31557**	-4.59560**
2016	AF	HA	271	-2.41763**	-2.40723**	-2.30610**	-2.56170**	-2.30610**	-6.94971**	-5.44187**
2016	SA	HA	439	-2.61249***	-2.58423***	-2.53118***	-2.56021***	-2.53408***	-6.76141**	-5.23596**
2015	NA	HA	427	-1.98721*	-1.94467*	-1.84345*	-2.09706**	-1.84345*	-6.05434**	-4.51307**
2015	EU	HA	598	-2.31515**	-2.29906**	-2.10967**	-2.55940**	-2.11243**	-9.60753**	-6.30093**
2015	AS	HA	781	-2.33782**	-2.24166**	-2.06491*	-2.43723***	-2.06719*	-9.13461**	-5.87052**
2015	AU	HA	42	-1.02395#	-1.05646#	-0.74531#	-1.56958#	-0.74531#	-2.00964#	-1.97120#
2015	AF	HA	60	-1.41022#	-1.41022#	-0.98168#	-1.96189*	-0.98168#	-3.27482*	-3.06385*
2015	SA	HA	51	-1.46910#	-1.49415#	-1.14704#	-2.17699**	-1.14704#	-3.61834**	-3.24343**
2009	NA	HA	1937	-2.71225***	-2.69035***	-2.68172***	-2.68177***	-2.68177***	-11.02073**	-6.53143**
2009	EU	HA	727	-2.75172***	-2.74101***	-2.71965***	-2.65838***	-2.72126***	-12.0238**	-7.66186**
2009	AS	HA	941	-2.76772***	-2.77568***	-2.74098***	-2.74463***	-2.74104***	-11.30094**	-7.06064**
2009	AU	HA	140	-2.60574***	-2.60355***	-2.57235***	-2.46645**	-2.57235***	-7.01505**	-5.837**
2009	AF	HA	124	-2.63355***	-2.59899***	-2.54884***	-2.46270**	-2.55424**	-7.27627**	-6.01422**
2009	SA	HA	66	-2.70238***	-2.71979***	-2.64890***	-2.53139***	-2.64890***	-5.1039**	-4.7318**

(TD = Tajima’s D)

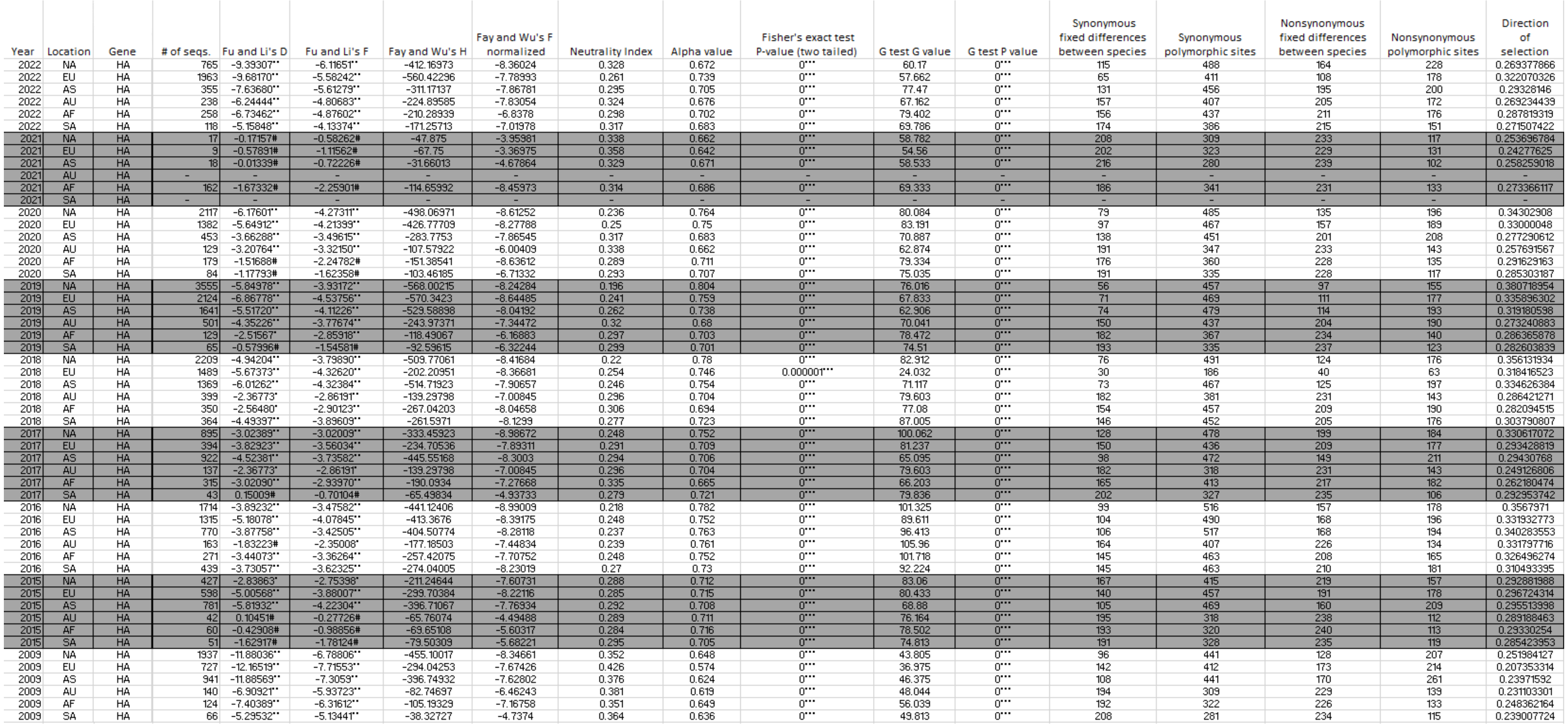
Supplement Table 4. Tajima’s D and Fu-Li’s D* and F* tests for neutrality selection for NA protein of 2009pdm influenza.



Year	Location	Gene	# of seqs.	TD	TD - Cod.	TD – Syn.	TD – Nonsyn.	TD - Silent	Fu-Li’s D*	Fu-Li's F*
2022	NA	NA	670	-2.52839***	-2.54079***	-2.49903***	-2.51488***	-2.50007***	-8.28129**	-5.73359**
2022	EU	NA	1657	-2.47577***	-2.48795***	-2.37566***	-2.62855***	-2.37779***	-8.57947**	-5.41552**
2022	AS	NA	336	-2.41006**	-2.38179**	-2.29701**	-2.43047***	-2.29997**	-7.72726**	-5.70097**
2022	AU	NA	243	-2.14433**	-2.14116**	-2.05656*	-2.20288**	-2.05656*	-6.94917**	-5.24634**
2022	AF	NA	215	-2.03407*	-1.97612*	-1.82504*	-2.15736**	-1.82504*	-5.78009*	-4.49688*
2022	SA	NA	88	-1.98964*	-1.97270*	-1.78112#	-2.12627*	-1.78112#	-5.49482*	-4.60784*
2021	NA	NA	17	-1.04478#	-1.09235#	-1.02695#	-1.17045#	-1.02695#	-1.88183#	-1.90061#
2021	EU	NA	16	-0.72430#	-0.73508#	-0.70739#	-0.36569#	-0.70739#	0.91902#	0.52004#
2021	AS	NA	10	-1.54137#	-1.54137#	-1.49441#	-1.44250#	-1.49441#	-1.90328#	-2.04579#
2021	AU	NA	-	-	-	-	-	-	-	-
2021	AF	NA	170	-1.30728#	-1.30728#	-1.02567#	-1.65445#	-1.02567#	-3.66582**	-3.07522**
2021	SA	NA	-	-	-	-	-	-	-	-
2020	NA	NA	2117	-2.47534***	-2.59376***	-2.52781***	-2.58248***	-2.52883***	-10.62478**	-6.20532**
2020	EU	NA	1455	-2.41748***	-2.40767***	-2.32180***	-2.45968***	-2.32349***	-8.78557**	-5.51902**
2020	AS	NA	565	-2.46142***	-2.47313***	-2.31959**	-2.58959***	-2.31959**	-8.40152**	-5.87300**
2020	AU	NA	140	-2.18886**	-2.19827**	-2.11150*	-2.16368**	-2.11150*	-4.76102**	-4.29335**
2020	AF	NA	198	-1.97254*	-1.98860*	-1.90077*	-2.04993*	-1.90077*	-5.24145**	-4.36587**
2020	SA	NA	94	-1.64389#	-1.68503#	-1.70505#	-1.48882#	-1.70505#	-2.94721*	-2.87221*
2019	NA	NA	3555	-2.29623***	-2.26078***	-2.11082***	-2.41417***	-2.11381***	-8.06289***	-4.81148**
2019	EU	NA	2281	-2.43658***	-2.42559***	-2.36961***	-2.38832***	-2.36951***	-8.47339**	-5.16256**
2019	AS	NA	1964	-2.50691***	-2.47963***	-2.40181***	-2.55608***	-2.40310***	-8.82405**	-5.41038**
2019	AU	NA	514	-2.22502**	-2.23886**	-2.16382**	-2.27629**	-2.16382**	-6.65018**	-4.89494**
2019	AF	NA	137	-2.19028**	-2.19921**	-2.24109**	-1.98084*	-2.24109**	-5.12677**	-4.52578**
2019	SA	NA	80	-1.82083*	-1.82083*	-1.89218*	-1.44488#	-1.89218*	-3.60602**	-3.44672**
2018	NA	NA	2355	-2.55018***	-2.52607***	-2.47703***	-2.46611***	-2.47703***	-8.59415**	-5.32484**
2018	EU	NA	1604	-2.49908***	-2.45457***	-2.41016***	-2.38969***	-2.41016***	-9.34905**	-5.73946**
2018	AS	NA	1597	-2.49822***	-2.48856***	-2.41807***	-2.48672***	-2.41807***	-7.50329**	-4.94882**
2018	AU	NA	80	-1.96475*	-1.89740*	-1.95175*	-1.63401*	-1.95175*	-4.00572**	-3.79860**
2018	AF	NA	378	-2.55086***	-2.52163***	-2.53034***	-2.34360**	-2.53250***	-7.52124**	-5.69766**
2018	SA	NA	75	-1.96396*	-.1.99061*	-1.95142*	-1.90012*	-1.95142*	-3.85858**	-3.70856**
2017	NA	NA	895	-2.40805***	-2.39650***	-2.31474**	-2.42883***	-2.31474**	-6.66906**	-4.81077**
2017	EU	NA	389	-2.47763***	-2.48134***	-2.50019***	-2.30521**	-2.49961***	-6.46175**	-5.08194**
2017	AS	NA	1015	-2.56995***	-2.56650***	-2.54748***	-2.48202***	-2.54863***	-8.17929**	-5.49563**
2017	AU	NA	140	-2.45712**	-2.47415**	-2.52041***	-2.24016**	-2.52041***	-6.82723**	-5.77272**
2017	AF	NA	346	-2.28905**	'-2.32630**	'-2.28987**	-2.24930**	-2.28987**	-6.71698**	-5.17467**
2017	SA	NA	46	-1.63252#	-1.70479#	-1.63484#	-1.53869#	-1.63484#	-1.83297#	-2.07805#
2016	NA	NA	1714	-2.67620***	-2.64213***	-2.59764***	-2.55142***	-2.59848***	-7.91718**	-5.28488**
2016	EU	NA	1401	-2.71726***	-2.69525***	-2.66854***	-2.61671***	-2.66355***	-11.04037**	-6.70684**
2016	AS	NA	891	-2.67139***	-2.66133***	-2.64154***	-2.58647***	-2.64134***	-8.80280**	-5.91366**
2016	AU	NA	154	-2.55829***	-2.55829***	-2.48502***	-2.54302***	-2.48502***	-6.55430**	-5.60631**
2016	AF	NA	279	-2.52727***	-2.51464***	-2.43155**	-2.54007***	-2.42524**	-7.31490**	-5.73695**
2016	SA	NA	75	-2.50352**	-.2.50352**	-2.49641**	-2.35410**	-2.49641**	-5.37115**	-5.04389**
2015	NA	NA	427	-2.25427**	-2.19594**	-2.18204**	-2.18204**	-2.18674**	-5.62516**	-4.47322**
2015	EU	NA	641	-2.46147***	-2.43781***	-2.42005***	-2.36964**	-2.42088***	-9.20171**	-6.20150**
2015	AS	NA	893	-2.49148***	-2.45070**	-2.38068**	-2.42656***	-2.38068**	-8.88416**	-5.81935**
2015	AU	NA	40	-1.49300#	-1.48359#	-1.48359#	-1.40345#	-1.48359#	-3.36492**	-3.21373**
2015	AF	NA	73	-1.41925#	-1.37644#	-1.36791#	-1.24303#	-1.36791#	-2.83213*	-2.71425*
2015	SA	NA	96	-1.18587#	-1.21270#	-1.32404#	-0.90299#	-1.32404#	-2.81406*	-2.54774*
2009	NA	NA	2090	-2.72679***	-2.69364***	-2.68779***	-2.55903***	-2.68779***	-11.82005**	-6.96205**
2009	EU	NA	771	-2.69809***	-2.70434***	-2.67299***	-2.62822***	-2.67169***	-10.39963**	-6.91355**
2009	AS	NA	980	-2.77093***	-2.75009***	-2.70699***	-2.70974***	-2.70699***	-12.73833**	-7.87646**
2009	AU	NA	140	-2.58497***	-2.58497***	-2.42854**	-2.62092***	-2.42854**	-6.74469**	-5.67407**
2009	AF	NA	131	-2.60824***	-2.60809***	-2.54443***	-2.49527***	-2.54443***	-5.66067**	-4.98431**
2009	SA	NA	64	-2.55804***	-2.55804***	-2.51425***	-2.25151**	-2.51425***	-4.08944**	-3.97858**

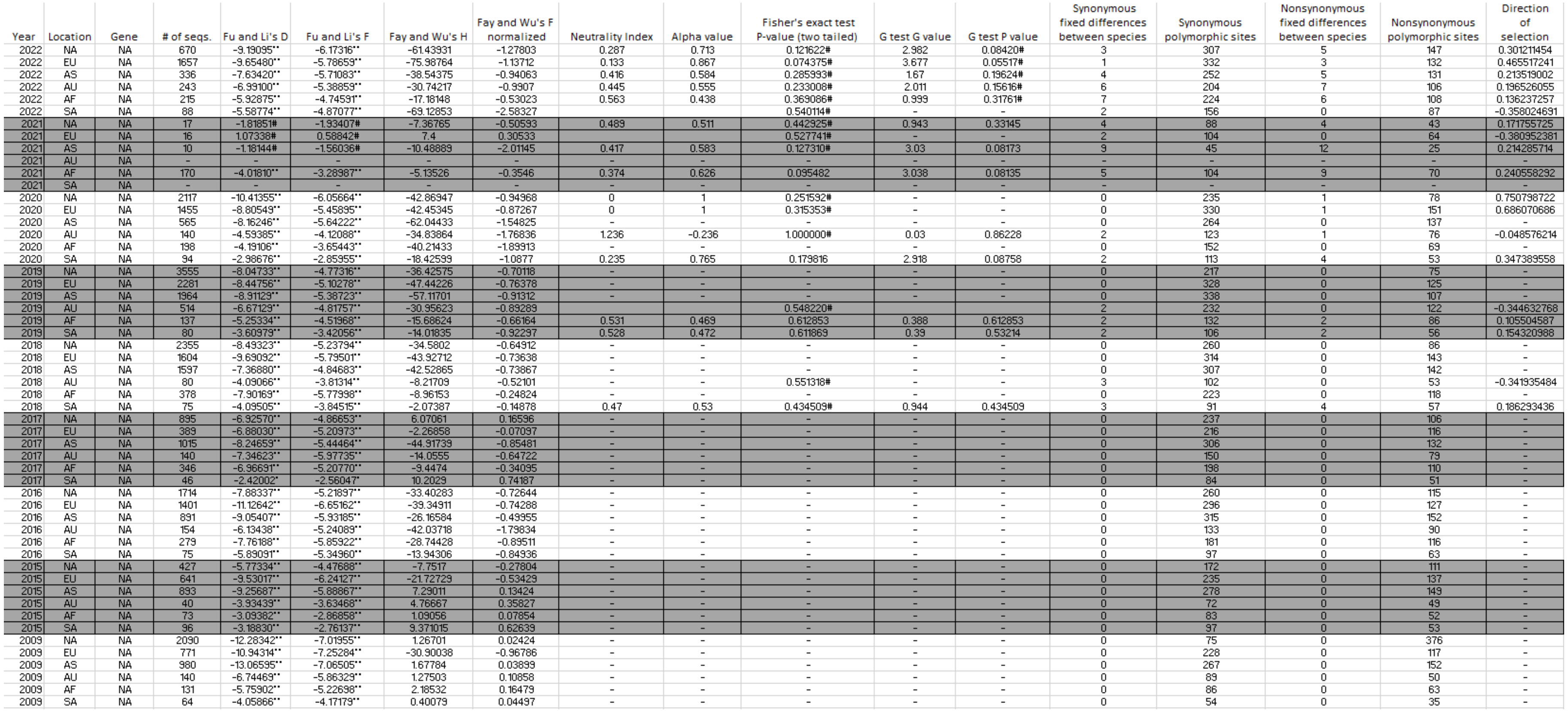
(TD = Tajima’s D)

Supplement Table 5. McDonald-Kreitman and Fu-Li's F and D tests with an outgroup for HA protein of 2009pdm influenza



Year	Location	Gene	# of seqs.	Fu and Li's D	Fu and Li's F	Fay and Wu's H	Fay and Wu's F normalized	Neutrality Index	Alpha value	Fisher's exact test P-value (two tailed)	G test G value	G test P value	Synonymous fixed differences between species	Synonymous polymorphic sites	Nonsynonymous fixed differences between species	Nonsynonymous polymorphic sites	Direction of selection
2022	NA	HA	765	-9.39307**	-6.11651**	-412.16973	-8.36024	0.328	0.672	0***	60.17	0***	115	488	164	228	0.269377866
2022	EU	HA	1963	-9.68170**	-5.58242**	-560.42296	-7.78993	0.261	0.739	0***	57.662	0***	65	411	108	178	0.322070326
2022	AS	HA	355	-7.63680**	-5.61279**	-311.17137	-7.86781	0.295	0.705	0***	77.47	0***	131	456	195	200	0.29328146
2022	AU	HA	238	-6.24444**	-4.80683**	-224.89585	-7.83054	0.324	0.676	0***	67.162	0***	157	407	205	172	0.269234439
2022	AF	HA	258	-6.73462**	-4.87602**	-210.28939	-6.8378	0.298	0.702	0***	79.402	0***	156	437	211	176	0.287819319
2022	SA	HA	118	-5.15848**	-4.13374**	-171.25713	-7.01978	0.317	0.683	0***	69.786	0***	174	386	215	151	0.271507422
2021	NA	HA	17	-0.17157#	-0.58262#	-47.875	-3.95981	0.338	0.662	0***	58.782	0***	208	309	233	117	0.253696784
2021	EU	HA	9	-0.57891#	-1.11562#	-67.75	-3.36975	0.358	0.642	0***	54.56	0***	202	323	229	131	0.24277625
2021	AS	HA	18	-0.01339#	-0.72226#	-31.66013	-4.67864	0.329	0.671	0***	58.533	0***	216	280	239	102	0.258259018
2021	AU	HA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2021	AF	HA	162	-1.67332#	-2.25901#	-114.65992	-8.45973	0.314	0.686	0***	69.333	0***	186	341	231	133	0.273366117
2021	SA	HA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2020	NA	HA	2117	-6.17601**	-4.27311**	-498.06971	-8.61252	0.236	0.764	0***	80.084	0***	79	485	135	196	0.34302908
2020	EU	HA	1382	-5.64912**	-4.21399**	-426.77709	-8.27788	0.25	0.75	0***	83.191	0***	97	467	157	189	0.33000048
2020	AS	HA	453	-3.66288**	-3.49615**	-283.7753	-7.86545	0.317	0.683	0***	70.887	0***	138	451	201	208	0.277290612
2020	AU	HA	129	-3.20764**	-3.32150**	-107.57922	-6.00409	0.338	0.662	0***	62.874	0***	191	347	233	143	0.257691567
2020	AF	HA	179	-1.51688#	-2.24782#	-151.38541	-8.63612	0.289	0.711	0***	79.334	0***	176	360	228	135	0.291629163
2020	SA	HA	84	-1.17793#	-1.62358#	-103.46185	-6.71332	0.293	0.707	0***	75.035	0***	191	335	228	117	0.285303187
2019	NA	HA	3555	-5.84978**	-3.93172**	-568.00215	-8.24284	0.196	0.804	0***	76.016	0***	56	457	97	155	0.380718954
2019	EU	HA	2124	-6.86778**	-4.53756**	-570.3423	-8.64485	0.241	0.759	0***	67.833	0***	71	469	111	177	0.335896302
2019	AS	HA	1641	-5.51720**	-4.11226**	-529.58898	-8.04192	0.262	0.738	0***	62.906	0***	74	479	114	193	0.319180598
2019	AU	HA	501	-4.35226**	-3.77674**	-243.97371	-7.34472	0.32	0.68	0***	70.041	0***	150	437	204	190	0.273240883
2019	AF	HA	129	-2.51567*	-2.85918**	-118.49067	-6.16883	0.297	0.703	0***	78.472	0***	182	367	234	140	0.286365878
2019	SA	HA	65	-0.57996#	-1.54581#	-92.59615	-6.32244	0.299	0.701	0***	74.51	0***	193	335	237	123	0.282603839
2018	NA	HA	2209	-4.94204**	-3.79890**	-509.77061	-8.41684	0.22	0.78	0***	82.912	0***	76	491	124	176	0.356131934
2018	EU	HA	1489	-5.67373**	-4.32620**	-202.20951	-8.36681	0.254	0.746	0.000001***	24.032	0***	30	186	40	63	0.318416523
2018	AS	HA	1369	-6.01262**	-4.32384**	-514.71923	-7.90657	0.246	0.754	0***	71.117	0***	73	467	125	197	0.334626384
2018	AU	HA	399	-2.36773*	-2.86191**	-139.29798	-7.00845	0.296	0.704	0***	79.603	0***	182	381	231	143	0.286421271
2018	AF	HA	350	-2.56480*	-2.90123**	-267.04203	-8.04658	0.306	0.694	0***	77.08	0***	154	457	209	190	0.282094515
2018	SA	HA	364	-4.49397**	-3.89609**	-261.5971	-8.1299	0.277	0.723	0***	87.005	0***	146	452	205	176	0.303790807
2017	NA	HA	895	-3.02389**	-3.02009**	-333.45923	-8.98672	0.248	0.752	0***	100.062	0***	128	478	199	184	0.330617072
2017	EU	HA	394	-3.82923**	-3.56034**	-234.70536	-7.89311	0.291	0.709	0***	81.237	0***	150	436	209	177	0.293428819
2017	AS	HA	922	-4.52381**	-3.73582**	-445.55168	-8.3003	0.294	0.706	0***	65.095	0***	98	472	149	211	0.29430768
2017	AU	HA	137	-2.36773*	-2.86191*	-139.29798	-7.00845	0.296	0.704	0***	79.603	0***	182	318	231	143	0.249126806
2017	AF	HA	315	-3.02090**	-2.93970**	-190.0934	-7.27668	0.335	0.665	0***	66.203	0***	165	413	217	182	0.262180474
2017	SA	HA	43	0.15009#	-0.70104#	-65.49834	-4.93733	0.279	0.721	0***	79.836	0***	202	327	235	106	0.292953742
2016	NA	HA	1714	-3.89232**	-3.47582**	-441.12406	-8.99009	0.218	0.782	0***	101.325	0***	99	516	157	178	0.3567971
2016	EU	HA	1315	-5.18078**	-4.07845**	-413.3676	-8.39175	0.248	0.752	0***	89.611	0***	104	490	168	196	0.331932773
2016	AS	HA	770	-3.87758**	-3.42505**	-404.50774	-8.28118	0.237	0.763	0***	96.413	0***	106	517	168	194	0.340283553
2016	AU	HA	163	-1.83223#	-2.35008*	-177.18503	-7.44834	0.239	0.761	0***	105.96	0***	164	407	226	134	0.331797716
2016	AF	HA	271	-3.44073**	-3.36264**	-257.42075	-7.70752	0.248	0.752	0***	101.718	0***	145	463	208	165	0.326496274
2016	SA	HA	439	-3.73057**	-3.62325**	-274.04005	-8.23019	0.27	0.73	0***	92.224	0***	145	463	210	181	0.310493395
2015	NA	HA	427	-2.83863*	-2.75398*	-211.24644	-7.60731	0.288	0.712	0***	83.06	0***	167	415	219	157	0.292881988
2015	EU	HA	598	-5.00568**	-3.88007**	-299.70384	-8.22116	0.285	0.715	0***	80.433	0***	140	457	191	178	0.296724314
2015	AS	HA	781	-5.81932**	-4.22304**	-396.71067	-7.76934	0.292	0.708	0***	68.88	0***	105	469	160	209	0.295513998
2015	AU	HA	42	0.10451#	-0.27726#	-65.76074	-4.49488	0.289	0.711	0***	76.164	0***	195	318	238	112	0.289188463
2015	AF	HA	60	-0.42908#	-0.98856#	-69.65108	-5.60317	0.284	0.716	0***	78.502	0***	193	320	240	113	0.29330254
2015	SA	HA	51	-1.62917#	-1.78124#	-79.50309	-5.68221	0.295	0.705	0***	74.813	0***	191	328	235	119	0.285423953
2009	NA	HA	1937	-11.88036**	-6.78806**	-455.10017	-8.34661	0.352	0.648	0***	43.805	0***	96	441	128	207	0.251984127
2009	EU	HA	727	-12.16519**	-7.71553**	-294.04253	-7.67426	0.426	0.574	0***	36.975	0***	142	412	173	214	0.207353314
2009	AS	HA	941	-11.88569**	-7.3059**	-396.74932	-7.62802	0.376	0.624	0***	46.375	0***	108	441	170	261	0.23971592
2009	AU	HA	140	-6.90921**	-5.93723**	-82.74697	-6.46243	0.381	0.619	0***	48.044	0***	194	309	229	139	0.231103301
2009	AF	HA	124	-7.40389**	-6.31612**	-105.19329	-7.16758	0.351	0.649	0***	56.039	0***	192	322	226	133	0.248362164
2009	SA	HA	66	-5.29532**	-5.13441**	-38.32727	-4.7374	0.364	0.636	0***	49.813	0***	208	281	234	115	0.239007724

Supplement Table 6. Population genetics analysis for nucleotide diversity for NA protein of 2009pdm influenza.



Year	Location	Gene	# of seqs.	Fu and Li's D	Fu and Li's F	Fay and Wu's H	Fay and Wu's F normalized	Neutrality Index	Alpha value	Fisher's exact test P-value (two tailed)	G test G value	G test P value	Synonymous fixed differences between species	Synonymous polymorphic sites	Nonsynonymous fixed differences between species	Nonsynonymous polymorphic sites	Direction of selection
2022	NA	NA	670	-9.19095**	-6.17316**	-61.43931	-1.27803	0.287	0.713	0.121622#	2.982	0.08420#	3	307	5	147	0.301211454
2022	EU	NA	1657	-9.65480**	-5.78659**	-75.98764	-1.13712	0.133	0.867	0.074375#	3.677	0.05517#	1	332	3	132	0.465517241
2022	AS	NA	336	-7.63420**	-5.71083**	-38.54375	-0.94063	0.416	0.584	0.285993#	1.67	0.19624#	4	252	5	131	0.213519002
2022	AU	NA	243	-6.99100**	-5.38859**	-30.74217	-0.9907	0.445	0.555	0.233008#	2.011	0.15616#	6	204	7	106	0.196526055
2022	AF	NA	215	-5.92875**	-4.74591**	-17.18148	-0.53023	0.563	0.438	0.369086#	0.999	0.31761#	7	224	6	108	0.136237257
2022	SA	NA	88	-5.58774**	-4.87077**	-69.12853	-2.58327			0.540114#	-	-	2	156	0	87	-0.358024691
2021	NA	NA	17	-1.81851#	-1.93407#	-7.36765	-0.50593	0.489	0.511	0.442925#	0.943	0.33145	4	88	4	43	0.171755725
2021	EU	NA	16	1.07338#	0.58842#	7.4	0.30533			0.527741#	-	-	2	104	0	64	-0.380952381
2021	AS	NA	10	-1.18144#	-1.56036#	-10.48889	-2.01145	0.417	0.583	0.127310#	3.03	0.08173	9	45	12	25	0.214285714
2021	AU	NA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2021	AF	NA	170	-4.01810**	-3.28987**	-5.13526	-0.3546	0.374	0.626	0.095482	3.038	0.08135	5	104	9	70	0.240558292
2021	SA	NA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2020	NA	NA	2117	-10.41355**	-6.05664**	-42.86947	-0.94968	0	1	0.251592#	-	-	0	235	1	78	0.750798722
2020	EU	NA	1455	-8.80549**	-5.45895**	-42.45345	-0.87267	0	1	0.315353#	-	-	0	330	1	151	0.686070686
2020	AS	NA	565	-8.16246**	-5.64222**	-62.04433	-1.54825	-	-	-	-	-	0	264	0	137	-
2020	AU	NA	140	-4.59385**	-4.12088**	-34.83864	-1.76836	1.236	-0.236	1.000000#	0.03	0.86228	2	123	1	76	-0.048576214
2020	AF	NA	198	-4.19106**	-3.65443**	-40.21433	-1.89913	-	-	-	-	-	0	152	0	69	-
2020	SA	NA	94	-2.98676**	-2.85955**	-18.42599	-1.0877	0.235	0.765	0.179816	2.918	0.08758	2	113	4	53	0.347389558
2019	NA	NA	3555	-8.04733**	-4.77316**	-36.42575	-0.70118	-	-	-	-	-	0	217	0	75	-
2019	EU	NA	2281	-8.44756**	-5.10278**	-47.44226	-0.76378	-	-	-	-	-	0	328	0	125	-
2019	AS	NA	1964	-8.91129**	-5.38723**	-57.11701	-0.91312	-	-	-	-	-	0	338	0	107	-
2019	AU	NA	514	-6.67129**	-4.81757**	-30.95623	-0.89289			0.548220#			2	232	0	122	-0.344632768
2019	AF	NA	137	-5.25334**	-4.51968**	-15.68624	-0.66164	0.531	0.469	0.612853	0.388	0.612853	2	132	2	86	0.105504587
2019	SA	NA	80	-3.60979**	-3.42056**	-14.01835	-0.92297	0.528	0.472	0.611869	0.39	0.53214	2	106	2	56	0.154320988
2018	NA	NA	2355	-8.49323**	-5.23794**	-34.5802	-0.64912	-	-	-	-	-	0	260	0	86	-
2018	EU	NA	1604	-9.69092**	-5.79501**	-43.92712	-0.73638	-	-	-	-	-	0	314	0	143	-
2018	AS	NA	1597	-7.36880**	-4.84683**	-42.52865	-0.73867	-	-	-	-	-	0	307	0	142	-
2018	AU	NA	80	-4.09066**	-3.81314**	-8.21709	-0.52101	-	-	0.551318#	-	-	3	102	0	53	-0.341935484
2018	AF	NA	378	-7.90169**	-5.77998**	-8.96153	-0.24824	-	-	-	-	-	0	223	0	118	-
2018	SA	NA	75	-4.09505**	-3.84515**	-2.07387	-0.14878	0.47	0.53	0.434509#	0.944	0.434509	3	91	4	57	0.186293436
2017	NA	NA	895	-6.92570**	-4.86653**	6.07061	0.16596	-	-	-	-	-	0	237	0	106	-
2017	EU	NA	389	-6.88030**	-5.20973**	-2.26858	-0.07097	-	-	-	-	-	0	216	0	116	-
2017	AS	NA	1015	-8.24659**	-5.44464**	-44.91739	-0.85481	-	-	-	-	-	0	306	0	132	-
2017	AU	NA	140	-7.34623**	-5.97735**	-14.0555	-0.64722	-	-	-	-	-	0	150	0	79	-
2017	AF	NA	346	-6.96691**	-5.20770**	-9.4474	-0.34095	-	-	-	-	-	0	198	0	110	-
2017	SA	NA	46	-2.42002*	-2.56047*	10.2029	0.74187	-	-	-	-	-	0	84	0	51	-
2016	NA	NA	1714	-7.88337**	-5.21897**	-33.40283	-0.72644	-	-	-	-	-	0	260	0	115	-
2016	EU	NA	1401	-11.12642**	-6.65162**	-39.34911	-0.74288	-	-	-	-	-	0	296	0	127	-
2016	AS	NA	891	-9.05407**	-5.93185**	-26.16584	-0.49955	-	-	-	-	-	0	315	0	152	-
2016	AU	NA	154	-6.13438**	-5.24089**	-42.03718	-1.79834	-	-	-	-	-	0	133	0	90	-
2016	AF	NA	279	-7.76188**	-5.85922**	-28.74428	-0.89511	-	-	-	-	-	0	181	0	116	-
2016	SA	NA	75	-5.89091**	-5.34960**	-13.94306	-0.84936	-	-	-	-	-	0	97	0	63	-
2015	NA	NA	427	-5.77334**	-4.47688**	-7.7517	-0.27804	-	-	-	-	-	0	172	0	111	-
2015	EU	NA	641	-9.53017**	-6.24127**	-21.72729	-0.53429	-	-	-	-	-	0	235	0	137	-
2015	AS	NA	893	-9.25687**	-5.88867**	7.29011	0.13424	-	-	-	-	-	0	278	0	149	-
2015	AU	NA	40	-3.93439**	-3.63468**	4.76667	0.35827	-	-	-	-	-	0	72	0	49	-
2015	AF	NA	73	-3.09382**	-2.86858**	1.09056	0.07854	-	-	-	-	-	0	83	0	52	-
2015	SA	NA	96	-3.18830**	-2.76137**	9.371015	0.62639	-	-	-	-	-	0	97	0	53	-
2009	NA	NA	2090	-12.28342**	-7.01955**	1.26701	0.02424	-	-	-	-	-	0	75	0	376	-
2009	EU	NA	771	-10.94314**	-7.25284**	-30.90038	-0.96786	-	-	-	-	-	0	228	0	117	-
2009	AS	NA	980	-13.06595**	-7.06505**	1.67784	0.03899	-	-	-	-	-	0	267	0	152	-
2009	AU	NA	140	-6.74469**	-5.86329**	1.27503	0.10858	-	-	-	-	-	0	89	0	50	-
2009	AF	NA	131	-5.75902**	-5.22698**	2.18532	0.16479	-	-	-	-	-	0	86	0	63	-
2009	SA	NA	64	-4.05866**	-4.17179**	0.40079	0.04497	-	-	-	-	-	0	54	0	35	-